Multilevel Models: A Tutorial on Applications and Uses in Human Factors Research

Courtney M. Goodridge[1, 2*], Jac Billington[1], Gustav Markkula[2], Richard M. Wilkie[1]

[1]School of Psychology, University of Leeds

[2]Institute for Transport Studies, University of Leeds

*c.m.goodridge@leeds.ac.uk

Abstract

Human behaviour is heterogeneous, a feature that is particularly important in Human Factors (HF) research where performance variability can have safety-critical implications. However, HF studies often focus on average effects between experimental conditions, treating individual differences as statistical noise rather than as meaningful information. Modelling behavioural heterogeneity can improve understanding of how systems affect users and support stronger theoretical development. Multilevel Models (MLMs) provide a flexible statistical framework for analysing hierarchical data structures, such as repeated measurements nested within individuals. Advances in statistical software have increased the accessibility of MLMs, yet many HF studies do not fully exploit their ability to model individual differences in experimental effects. One barrier is limited availability of practical tutorials demonstrating how MLMs can be applied to typical HF datasets. This manuscript addresses this gap by providing a practical introduction to MLMs for HF researchers. First, we review MLM principles and their relevance to HF research. We then present two worked examples. Study 1 demonstrates MLMs as an extension of linear regression for continuous predictors; Study 2 applies MLMs to factorial designs and discusses strategies for managing convergence in complex random-effects structures. Analyses are supported by example datasets and code in R, Python, and MATLAB.

Keywords: Multilevel Models; Mixed-Effects Models; Human Factors; Takeovers; Mental Workload

1 Introduction

1.1 Background

Human behaviour is inherently variable. Even under controlled experimental conditions, where the same person responds to the same stimuli in consecutive trials, their responses (e.g., a reaction time) will not be identical. Furthermore, two different people with, albeit very similar demographics, are even more unlikely to respond in the same way to identical stimuli. Despite this observation, it is typical in experimental research to focus on the average change between conditions when generating inferences from data. Human Factors (HF) research is wide-ranging (e.g., road/air/rail transport, plant operations, defence, etc.); therefore, statistics can be used to generate a variety of inferences to answer research questions. The stated applied motivation for statistical analyses in HF is typically one of the following: to guide the design of technology (with respect to, e.g., system features and functionality, user interfaces, and operator state estimation) or to guide regulations, procedures, and training.

The stated motivation for the analysis can also be more scientifically oriented, e.g., to *understand* or *model* aspects of human technology use, but if so, this is typically justified with reference to the intended applications listed above. For all these applications, it is useful to know the average main effects and interactions of the independent variables. However, an analysis of *only* main effects and interactions rarely tells the full story, as it is unusual for the variability in human response to be attributed primarily to variability in experimental procedures or measurement (treatment or measurement error; Bolger et al., 2019). In practice, there are virtually always differences within and between participants in their response to experimental manipulations, so a restrictive focus on just the central tendency presents a lost opportunity for developing theories and understanding the involved behavioural phenomena.

There are many disciplines within HF where human response heterogeneity is deemed vital for answering research questions. Firstly, consider the range of studies that have investigated how

individual differences (i.e., differences associated with personality, motivation, cognitive abilities, etc) impact a range of HF topics including aviation (Grassmann et al. 2017) automation (Jipp 2014; Lin 2017), robotics, (Chen et al. 2013), artificial intelligence (Matthews et al. 2021), and automated and manual driving (Körber and Bengler 2014; Lansdown 2002; Wiebel-Herboth et al. 2021). Some of these studies rely on relatively simplistic approaches to assessing heterogeneity, such as dividing participants into groups based on trait levels or experience (e.g., low versus high trait scores, or novice versus experienced users) and comparing mean differences in dependent variables (Chen et al. 2013; Grassmann et al. 2017; Lansdown 2002; Lin 2017). However, there are also strong examples in the literature demonstrating the benefits of moving beyond central tendency (Clegg 2000; Jipp 2014; Loske and Klumpp 2022; Matusiak et al. 2017; Wiebel-Herboth et al. 2021). These studies typically employ Multilevel Models (MLMs) which enable researchers to estimate the average effect of an experimental manipulation *and* to capture how these effects vary across individuals within the sample.

1.2 What are Multilevel Models?

Multilevel models (MLMs) are statistical models designed for hierarchical, clustered, or nested data structures where observations are grouped at higher levels (Singmann and Kellen 2019). The term "Multilevel Model" is used throughout this manuscript, however, the terminology surrounding MLMs is varied. Multilevel models, mixed effects models, random effects models, random-coefficient models, and hierarchical models can all refer to the same analytic framework (Shin 2009). MLMs assess variability attributed to the experimental manipulation alongside random variability via the inclusion of fixed and random effects, respectively (Harrison et al. 2018). Typically, fixed effects represent the effects that are observed experimentally. They are generally brought about by manipulating independent variables which are hypothesised to generate systematic variability in the observed response. Each level of the

fixed effect represents a particular condition or contrast, and the MLM aims to predict the mean of condition responses (Lo and Andrews 2015). Conversely, random effects are often represented as a grouping variable (e.g., grouping responses from one individual).

Because random effects are grouping structures, they are inherently discrete and thus continuous variables cannot be included as random effects (Winter 2019). A typical example of a random effect, and one relevant to HF, are participants within a repeated measures experimental design. The sample of participants used in an experiment are merely a subgroup of people from the wider population. Each participant provides multiple responses across all available conditions. The inclusion of random effects allows for the estimation of variability at two levels: within individual participants and between different participants. Random effects need not always be participants; any discrete grouping structure that has multiple observations can act as a random effect – for example, timepoints at which participants are tested during a longitudinal study (Grech et al. 2009) or even individual studies akin to a meta-analysis (Zaal et al. 2021). Conversely, fixed effects focus on systematic variability that is explicitly investigated (i.e., mean differences between condition levels). A common confusion with fixed and random effects is what makes them "fixed" and "random". Fixed effects are considered "fixed" because they are modelling the effect of a predictor variable that is common across participants within the sample (Brown 2021). Conversely, whilst random effects are consistent within an individual (e.g., one participant will be systematically different from another), the ultimate source of this variability is unknown and is considered "random" (Brown 2021).

1.3 Examples of MLMs in HF

There are several concrete examples of MLM use within HF research. In intralogistics, studies of forklift-based order picking have shown that variability in task completion time is not simply random noise but is systematically attributable to individual operators (Clegg 2000; Loske and Klumpp 2022; Matusiak et al. 2017). For example, 10.3% of the variance in picking time is

explained by stable individual differences when using MLMs (Matusiak et al. 2017); similarly, 15.1% of variance in batch execution time is attributable to differences between operators (Loske and Klumpp 2022). These findings illustrate what would be missed using traditional approaches such as analysis of variance (ANOVA). A group-based analysis (e.g., comparing mean performance between “fast” and “slow” operators, or between experimental conditions) would collapse this variability into residual error, obscuring the fact that a substantial proportion of performance differences is consistently tied to individuals. In contrast, MLMs explicitly partition variance into within- and between-participant components, revealing that operator-specific effects are both measurable and practically meaningful. This is not merely about statistically controlling for individual differences; rather, it involves quantifying and interpreting them as a key source of system variability. In the context of order picking, individual differences may reflect factors such as spatial strategies, risk tolerance, experience, or interaction styles with the equipment. Because these differences account for a non-trivial proportion of batch execution time, they have direct operational implications. Modelling heterogeneity can inform human-centred interventions, such as adaptive task allocation or incentive-based sociotechnical systems, which would not emerge from analyses focused solely on average performance (Loske and Klumpp 2022).

MLMs enable researchers to examine heterogeneity by determining whether the effects of experimental manipulations differ meaningfully across individuals. In doing so, they avoid reliance on data aggregation practices that can obscure data structure. For example, averaging across trials within participants can create unwarranted confidence by removing variability and masking the underlying number of observations contributing to each estimate (McElreath 2018). By modelling data at the appropriate level, MLMs allow HF researchers to estimate average effects and interactions with greater precision and statistical power than conventional approaches such as ANOVA.

There is a sense that researchers are becoming aware of MLMs, with the general domains of Psychological, Behavioural, and Cognitive Science, shifting from using Repeated Measures (RM) ANOVAs to implementing MLMs as the default analytic method (Boisgontier and Cheval 2016; Harrison et al. 2018; Haverkamp and Beauducel 2017; Lo and Andrews 2015; Muradoglu et al. 2023; Nalborczyk et al. 2019). One reason for the shift towards MLMs are developments in software that allows researchers to fit these models, whether using Frequentist (Bates et al. 2015) or Bayesian frameworks (Bürkner 2017). Improved software availability and the purported benefits of MLMs (Barr et al. 2013; Nalborczyk et al. 2019) has resulted in the increased usage of MLMs when analysing data generated from experiments investigating HF topics. A cursory search in the empirically oriented Human Factors journal reveals that '"mixed effects" models' were mentioned four times in manuscripts published between 2004-2005. A decade later, it was five (2014-2015); a further decade and this reaches 30 (2024-2025). For context, articles mentioning "repeated measures" and "ANOVA" totalled 27 between 2004-2005; 46 a decade later (2014-2015); and 56 a decade after that (2024-2025); suggesting that despite increasing MLM popularity, RM ANOVAs are still in the majority.

Recently, the mention of MLMs in the Human Factors journal has focused on a range of topics including manual and automated vehicles (Kamaraj et al. 2024; Zgonnikov et al. 2024), interactions with artificial intelligence (Dunning et al. 2024; Kim et al. 2025); and pedestrian behaviour (Malik et al. 2024, 2025). Despite increased usage, the interpretation of these models is still largely restricted to accounting for individual differences, (Kamaraj et al. 2024; Zgonnikov et al. 2024), increasing statistical power (Kim et al. 2025; Malik et al. 2025), or accounting for repeated measurements within a research design (Dunning et al. 2024; Kim et al. 2025). Whilst these interpretations are all valid, they represent a missed opportunity for further theoretical development in their respective fields by not delving deeper into the variance parameters that provide MLMs with their core functionality.

The rise in introductory tutorials explaining MLM theory and how MLMs address issues related to data aggregation and statistical power (Brown 2021; Singmann and Kellen 2019) are likely a contributing factor to (or a consequence of) increased MLM usage. The current tutorial complements these by providing an accessible entry point for readers who are new to the approach and work within HF - this will be demonstrated in Study 1 of this manuscript. In Study 2, we will extend beyond these foundations and provide insights to researchers who may already use MLMs within the field of HF. This manuscript goes beyond this prior work by placing more focus on how to investigate whether average effects are indicative of the overarching population. Exploring the *range* and *size* of the variability relative to the average person is implicit within the MLMs, yet it is rarely explicitly investigated in published manuscripts. This is particularly relevant for HF, given that safety critical human-system interactions are likely to be at the tail ends of human variabilities. Rather than using MLMs to model heterogeneity as a means to an end (i.e., to account for differences between participants to increase statistical power), this tutorial highlights that modelling heterogeneity is an end in and of itself (i.e., reporting the size, nature, and extent of variance in experimental effects within a population to enhance theoretical development).

1.4 Current work

This tutorial is aimed at HF researchers; those who are MLM novices and those who have some experience but would like to make fuller use of MLM capabilities to answer research questions about variability. Specific focus will be placed on human interactions with automated vehicles (AVs). Between manually driven vehicles (i.e., defined by the Society of Automotive Engineers [SAE] as Level 0 [L0]) and full AVs (i.e., L5) there will be intermediate levels of automation where the driver's role changes from vehicle operator to vehicle supervisor. L2 vehicles control the lateral and longitudinal dynamics of the vehicle, yet the driver must continuously monitor the road and system for hazards and must take back control if needed to maintain safety. The

use of MLMs for investigating driver behaviour is vital when it comes to evaluating safety. Despite purported safety benefits of automated driving functions (Gajera et al. 2022), there is still concern regarding how drivers take back manual control - especially when automated systems fail unexpectedly - and the ability of drivers to monitor the road and automated system efficiently whilst engaging in non-driving related tasks (NDRTs).

The datasets used in this tutorial focus on L2 driving. The dataset presented in Study 1 (N = 12) comes from an experiment investigating how drivers intervene from an L2 driving system that fails without warning. To provide a conceptual understanding of MLMs for a repeated measures design, these data will be first analysed using common techniques (i.e., linear regression). Following this, a step-by-step tutorial will explain how to use a MLM to account for individuals within the sample and estimate the mean for each participant rather than focusing only on the central tendency. In Study 2, a second dataset (N = 38) is introduced investigating the influence of increased mental workload (MWL) on eye movements during hands-off L2 driving. This study outlines how MLMs are used to analyse data from common experimental designs (i.e., a 2 x 2 factorial design) and how they can be used to provide inferences on the size and relevance of heterogeneity within a population. This section will specifically focus on why exploring the heterogeneity of effects is meaningful in and of itself.

2 Study 1: Predicting Takeover Response to Silent Automated Vehicle Failures

2.1 Experimental description

This study analysed driver responses to silent vehicle automation failures (i.e., system failures that do not alert the driver - similar to Louw et al., 2019; Mole et al., 2019). Participants were tasked with monitoring an L2 driving system and taking back control when necessary. The independent variable was failure severity and was operationalised via time to lane crossing at failure (TLC_F; i.e., the available time budget at the point of failure until the vehicle exceeded

the lane boundary if no steering input was made). TLC_F was manipulated across four levels - 2.23 s, 4.68 s, 7.12 s, and 9.55 s - with lower values producing more severe failures. The dependent variable was time to lane crossing at takeover (TLC_T; i.e., the amount of time it would take the vehicle to leave the road at the point of taking over control - Boer, 2016; Godthelp et al., 1984). This metric is often referred to as "safety margin" whereby lower values indicate less time before the driver exceeds the lane boundaries. In the original experiment, one driving block involved supervising the driving system; the other block involved completing a secondary task whilst supervising the driving system. Only data involving the block without the secondary task is used for this tutorial; once secondary task data were removed, the sample size for the following sections was N = 12 with each participant having six observations per TLC_F level.

2.2 Introducing MLMs

The typical way to understand how failure severity impacts the safety margin of drivers would be to model how driver safety margin changed as a function of the silent failure criticality, and one common way to achieve this would be to apply a RM ANOVA. However, to progress towards an MLM, this tutorial will start with a linear regression (RM ANOVAs are just special cases of linear regression where predictors are discrete rather than continuous - see Ben-Shachar, 2018). Equation 1 highlights the linear regression model equation. An ordinary regression model aims to predict the mean of a response variable Y_i via a linear combination of an intercept β_0 parameter, and a slope parameter β_1 which quantifies the influence of a predictor X_i:

$$Y_i \sim N(\mu_i, \sigma_e) \qquad 1$$

$$\mu_i = \beta_0 + \beta_1 X_i$$

The outcome Y_i is assumed to follow a normal distribution with mean μ_i and standard deviation σ_e. The mean is expressed as a linear function of predictor X_i, parameterised by the intercept β_0 and slope β_1. Here, subscript i denotes individual observations. The parameter σ_e represents the standard deviation of the residual errors, capturing variability in Y_i that is not explained by the linear relationship with X_i. If we apply the model in Equation 1 to the TLC_T data, the following model setup can be generated:

$$TLC_{T_i} \sim N(\mu_i, \sigma_e) \qquad 2$$

$$\mu_i = \beta_0 + \beta_F F_i$$

Where predictor F_i represents the failure severity level for observation i, treated as a continuous variable. A summary of the fitted model can be found in Table 1 (R/Python/MATLAB code for this and all analyses is available alongside supplementary material (https://osf.io/sruhn/overview)

Table 1: Summary of linear regression model for TLC_T

		TLC_T	
Predictors	*Estimates*	*CI*	*p*
β_0	.475	[.317, .633]	**<0.001**
β_F	.301	[.277, .326]	**<0.001**
Observations:	286		

The intercept (β_0) represents the mean value of Y_i when the predictors are zero[1]. The slope parameter (β_F) highlights the change in TLC_T for every 1 s increase in TLC_F and indicates a

[1] Note that while the intercept is needed to define the assumed linear relationship between predictor and response variables, it is not always necessarily interpretable in itself, e.g., in cases such as here where the predictor cannot

statistically significant ($p < 0.001$) change in TLC_T as a function of TLC_F. The parameter predicts a 0.301 s increase in safety margin for every 1 s increase in TLC_F. The 95% confidence intervals highlight that this effect could range from 0.277 s to 0.326 s. Overall this suggests that drivers respond with higher safety margins, which are associated with safer responses, when the criticality of the situation is reduced.

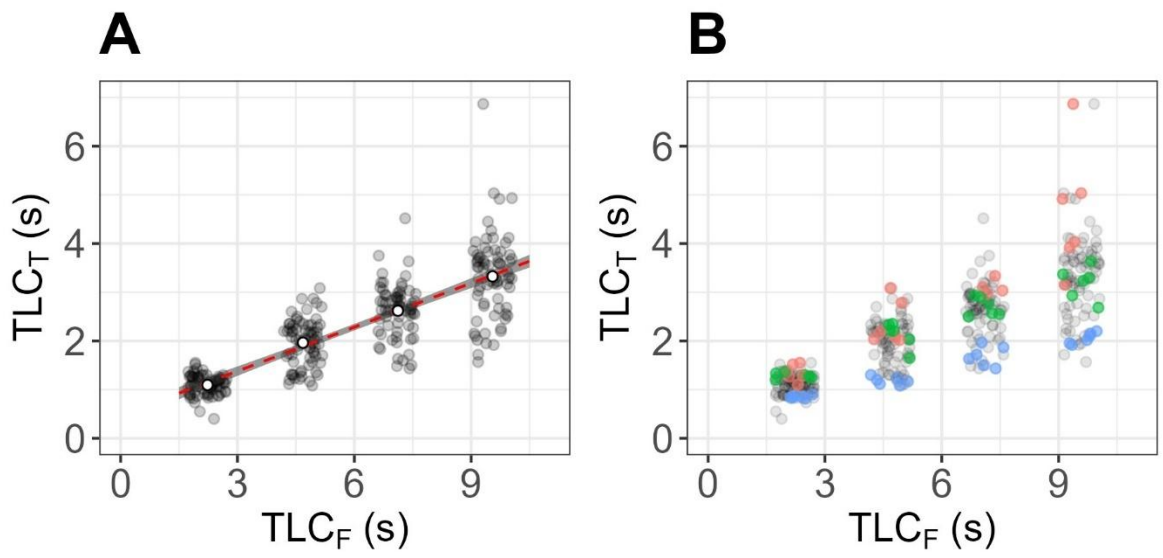

Figure 1: (A) Model parameters (red dashed line) plotted against sample means (open circles) over raw data. (B) The multilevel structure plotted over raw data. Each participant has multiple observations within each failure condition. Observations belonging to three example participants (blue, green and pink filled circles) appear clustered together.

However, there is a problem with the model specified in Equation 2; it fails to account for variation that, from a theoretical perspective, we know must exist. By only estimating a single intercept and slope across all participants, the model assumes that predicted TLC_T values within each TLC_F condition are identical for every individual. As outlined in the introduction, this assumption is implausible, as individuals differ systematically in their performance. These

meaningfully equal zero (since while the lane keeping system is still functioning, it would be expected to keep $TLC_F > 0$ at all times).

systematic differences are seen visually in Figure 1B where three participants are highlighted in different colours; observations pertaining to the same participant appear clustered together. Ignoring between-participant variability has important statistical consequences for the estimation of TLC_T. Most notably, the uncertainty (e.g., the standard error) surrounding the parameter of interest is underestimated. Without explicitly modelling participant-level differences, the model implicitly assumes that every trial - regardless of the participant from whom it was obtained - was generated by the same underlying process. In the literature, this is known a pseudo-replication; when a researcher assumes they are collecting independent samples, whereas some samples are not independent because they are collected from a single overriding group; in this case, that group is the participant (Davies and Gray 2015; Harrison et al. 2018). This assumption leads to artificially narrow confidence intervals, as illustrated in Figure 1A by the unusually tight 95% confidence bounds around the regression line.

Because the model does not incorporate participant-level variability, it is unable to appropriately calibrate the uncertainty surrounding the point estimate to reflect differences between individuals. By treating all observations as though they arise from a single, homogeneous data-generating process, the model assumes that new participants conform to the same underlying structure observed in the original data. When this assumption is violated - as is likely the case given inherent differences between people - the model's estimates of precision will be overly optimistic, and its apparent performance will not generalise accurately to new participants whose data-generating processes differ from the original sample. Overall, this highlights how linear regression is inappropriate to adequately model the dependencies amongst data points.

2.3 Varying intercept-fixed slope model

We established that the data-generating process is unlikely to be identical across individuals. Rather, each participant can be conceptualised as having their own underlying mechanism that generates observed responses. This naturally implies that data points arising from the same participant are more similar to one another than to data points generated by a different participant, because they share the same participant-specific process. In the context of transitions of control - when reacting as quickly as possible to silent failures - this means that some drivers will systematically produce longer reaction times, while others will consistently respond more quickly. These systematic differences may reflect stable characteristics of the individual data-generating process, such as differences in perceptual sensitivity, cognitive processing speed, or motor execution.

If each participant's responses are generated by a slightly different process, then a single global intercept is insufficient to capture this structure. An improvement over ordinary linear regression is allowing each participant to have their own intercept, representing their individual displacement from the overall population mean. This permits the entire linear relationship to shift upward or downward depending on the participant, while retaining a common slope across conditions. This is precisely what an MLM with varying intercepts and fixed slopes accomplishes. Such a model extends ordinary linear regression by explicitly acknowledging that observations are generated within higher-level units, and by assigning each member of the random-effects grouping variable (i.e., participant) a distinct intercept.

By doing so, the model recognises that responses within a participant share a common participant-specific data-generating process. It therefore captures systematic between-participant variability while simultaneously accounting for the non-independence of observations, because trials generated by the same participant are constrained to share the same

intercept (Royle 2013). Therefore, MLMs more accurately reflect the hierarchical structure of the data and provide appropriately calibrated estimates of uncertainty. A varying intercept model can be defined as follows:

$$TLC_{T_{ij}} \sim N(\mu_{ij}, \sigma_e) \qquad 3$$

$$\mu_{ij} = (\beta_0 + \beta_{0_j}) + \beta_F F_i$$

$$\beta_{0_j} \sim N(0, \sigma_{\beta_{0_j}})$$

In this model, as before, subscript i indexes observations, however j now indexes participants. The predictor F_i still represents the failure severity level for observation i, treated as a continuous variable. The mean (μ_{ij}) is modelled as a function of the population-level intercept (β_0), the fixed slope (β_F), and a participant-specific intercept (β_{0_j}). This participant-specific intercept allows the predicted mean to vary across participants, meaning that each participant has their own intercept while sharing a common slope for the predictor (see Figure 2). This accounts for the non-independence of observations within participants while also allowing the model to estimate how much participants differ in their predicted mean responses.

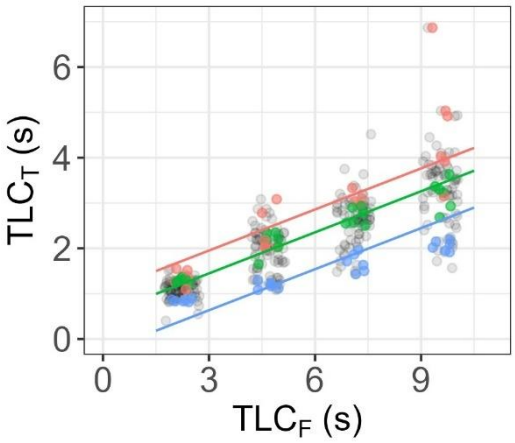

Figure 2: Allowing the intercepts to vary by participant allows the model to capture some of the variability between participants and constrains non-independent observations to the same intercept. The blue, green, and pink dots and lines represent the same subset of three example participants represented in Figure 1B.

This hierarchical structure reflects the assumption that participants are sampled from a broader population. Rather than estimating completely independent intercepts for each participant, the model assumes that individual intercepts represent deviations from the overall population intercept (β_0)[2]. These deviations are assumed to follow a zero-centred normal distribution with a standard deviation ($\sigma_{\beta_{0_j}}$). This parameter represents the variability of participant intercepts in the population and therefore quantifies the degree to which participants differ in their predicted mean. Estimating this parameter allows the model to generalise beyond the observed sample to the wider population from which participants were drawn (Barr et al. 2013).

In addition to between-participant variability, the model estimates the residual standard deviation (σ_e) as before, which captures the remaining variation in $\mathrm{TLC_T}$ across observations after accounting for the predictor F_i and participant-specific intercepts. This residual component represents within-participant variability across trials (Nalborczyk et al. 2019). A summary of the varying intercept model is presented in Table 2.

[2] This description reflects what is going on "under the hood" in these models. A combination of the MLM parameters and participant data can be used to generate predictions which are known as "conditional modes" (Kliegl et al. 2011). These effectively quantify how much each participant deviates away from the grand mean and can be used to estimate each participant's intercepts.

Table 2: Summary of fixed and random effects for varying intercept model for TLC_T

Fixed effects		TLC_T	
Predictors	*Estimates*	*CI*	*p*
β_0	.475	[.244, .707]	**<0.001**
β_F	.302	[.282, .322]	**<0.001**
Random Effects			
σ_e	.478		
$\sigma_{\beta_{0j}}$	.331		
N	12		
Observations:	286		

2.4 Varying intercept-varying slope model

An additional source of variability that MLMs can account for comes from the fact that a variable that is modelled as a fixed effect may actually have different influences on different participants. In the varying-intercept-fixed-slope model above, by keeping the effect of TLC_F fixed, there is an assumption that the extent to which participants are affected by *changing* failure severities is the same. However, this assumption might not hold. People may not only vary in how quickly they respond in general, but also in how sensitive they are to changes in failure severity. In terms of the model, this corresponds to allowing the slope to vary between different participants, i.e., making the change in safety margin for a unit increase in TLC_F participant-specific (Harrison et al. 2018). Indeed, from Figure 2 we can see that only varying the intercept may not have produced optimal fits. For example, the bottom, blue regression line entirely misses the data point cluster at $TLC_F \approx 3$ s. The fixed-slope model cannot account well for these data, because the regression line can only move up and down the y axis to produce the best fit. A model with a varying intercept *and* varying slope can be specified as follows:

$$TLC_{T_{ij}} \sim N(\mu_{ij}, \sigma_e) \quad (4)$$

$$\mu_{ij} = \left(\beta_0 + \beta_{0_j}\right) + \left(\beta_F F_i + \beta_{F_j} F_i\right)$$

$$\begin{bmatrix} \beta_{0_j} \\ \beta_{F_j} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, S\right)$$

$$S = \begin{pmatrix} \sigma^2_{\beta_{0_j}} & \rho_{\mathrm{F0}}\sigma_{\beta_{\mathrm{F}_j}}\sigma_{\beta_{0_j}} \\ \rho_{0F}\sigma_{\beta_{0_j}}\sigma_{\beta_{F_j}} & \sigma^2_{\beta_{F_j}} \end{pmatrix}$$

As in the previous models, subscript i indexes observations, j indexes participants, and the predictor F_i represents the failure severity level associated with observation i, treated as a continuous variable. The participant-specific intercepts and slopes are assumed to follow a multivariate normal (MVN) distribution with mean vector $[0, 0]^T$ and variance-covariance matrix S. This matrix describes the population distribution of the random effects. One component of this matrix is the variance of the random slopes, ($\sigma^2_{\beta_{F_j}}$), whose square root ($\sigma_{\beta_{F_j}}$) represents the standard deviation of participant-specific slopes in the population[3]. Estimating this parameter quantifies how much the effect of failure severity varies across participants. S also includes the covariance between the random intercepts and random slopes, typically expressed through the correlation parameter ($\rho_{0F}\sigma_{\beta_{0_j}}\sigma_{\beta_{F_j}}$). This term describes the extent to which participant-specific intercepts and slopes tend to co-vary. For example, it captures whether participants with higher intercepts also tend to have steeper or shallower slopes. In practice, intercepts and slopes are often assumed to exhibit some degree of covariation, which

[3] In Equation 4 the variability is expressed using the variance ($\sigma^2_{\beta_{0_j}}$ and $\sigma^2_{\beta_{F_j}}$) because the standard deviation of the random slopes ($\sigma_{\beta_{F_j}}$) and intercepts ($\sigma_{\beta_{0_j}}$) are multiplied by themselves in the covariance matrix generating a correlation of 1 for $\sigma^2_{\beta_{0_j}}$ and $\sigma^2_{\beta_{\mathrm{F}_j}}$ respectively (Nalborczyk et al. 2019). For reporting, the standard deviation will be used.

is why they are modelled jointly using a multivariate normal distribution rather than as independent normal distributions (Baayen 2004; Barr et al. 2013; Davidian and Giltinan 2003; Stawski 2013). The summary of the fitted model corresponding to Equation 4 is presented in Table 3.

Table 3: Summary of fixed and random effects for varying intercept-varying slope model for TLC_T

Fixed effects		TLC_T	
Predictors	*Estimates*	*CI*	*p*
β_0	.475	[.346, .604]	**<0.001**
β_F	.302	[.257, .347]	**<0.001**
Random Effects			
σ_e	.442		
$\sigma_{\beta_{0_j}}$	.086		
$\sigma_{\beta_{F_j}}$	.068		
$\rho\sigma_{\beta_{0_j}}\sigma_{\beta_{F_j}}$	-.820		
N	12		
Observations	286		

Once again, the fixed effects in this model are similar to those in the previous model; the differences come when looking at the random effects structure. The first thing to note in this new model is the reduction in individual intercept variability. For the varying intercept-fixed slope model, $\sigma_{\beta_{0_j}}$ was estimated as .331 s; for the varying intercept-varying slope model, $\sigma_{\beta_{0_j}}$ is estimated as .086 s. For the varying intercept-fixed slope model, if a participant responded with a high safety margin for less severe failures, the model could only increase the intercept. Consequently, this artificially inflated the estimated intercept variation. Conversely, the varying intercept-varying slope model allows the *slope* of the regression line to change; if drivers

respond with higher safety margins for less severe failures, the intercept need not change whereas the slope can change to produce the best fit. Another thing to note is the change is the residual error (σ_e). This has reduced from .478 s in the varying intercept model to .442 s in the varying intercept-varying slope model. Residual error is a measure of the deviation of data points around a specific participant's regression line; the varying intercept-varying slope model has a lower residual error because each participant's slope is better tailored to fit the individual data points with the addition of the random slope parameter (Brown 2021).

The first of the new parameters is the standard deviation of the random slopes ($\sigma_{\beta_{F_j}}$ = .068) which indicate that participant slopes varied around the average slope by around .07 s. A participant one standard deviation below the average would expect to have an increase in safety margin of .233 s for every 1 s increase in time budget; a participant one standard deviation above the average would be expected to have an increase in safety margin of .370 s for every 1 s increase in time budget. Allowing slopes to vary between different participant allows a HF scientist to explicitly model how different participants change their safety margin as a function of silent failure criticality (see Figure 3).

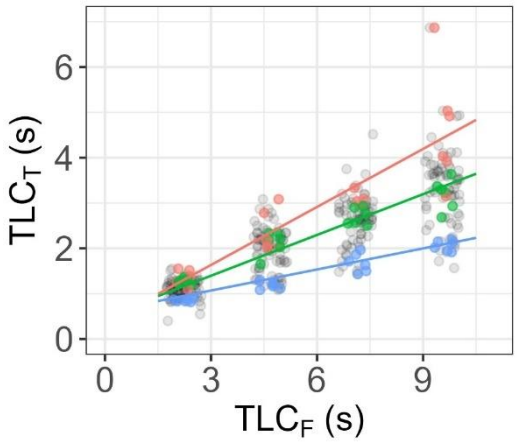

Figure 3: Allowing the slopes and intercepts produces a better fit for each participant regarding their TLC_T for a given level of TLC_F. Blue, green, and pink circles and lines are the same participants highlighted in Figure 2

and Figure 1B. The intercepts are very similar, whereas the sensitivity to the manipulation of failure severity changes depending on the participant producing a better fit for the highlighted individuals.

2.5 Conclusion

This study has provided a first example of how to use MLMs to understand a HF dataset. Ordinary linear regression was introduced to explain why it was inappropriate for data obtained via a repeated measures design. Random intercepts and slopes demonstrated how MLMs deal with violations of independence; by constraining individual data points belonging to the same person to the same regression line. This increases the statistical power and allows researchers to estimate individual level effects for an independent variable of interest. By modelling the full distribution of individual effects rather than relying on a single average estimate, MLMs reveal the range of plausible outcomes across the population - information that is invisible within ordinary linear regression. In safety-critical HF contexts, this matters considerably: a policy calibrated to the average driver response time may be dangerously inadequate for the slowest-responding individuals in the population, a vulnerability that would remain entirely hidden had variance across individuals been absorbed into residual error rather than explicitly estimated. This example merely scratches the surface of how MLMs can be used; in the supplemental material, further examples are provided that extend Study 1 to investigate how to quantitatively estimate how similar or different groups are within the random effects structure. This is akin to assessing to the intralogistics example provided in the introduction (Clegg 2000; Loske and Klumpp 2022; Matusiak et al. 2017) allowing for the assessment of how much of the variance in safety margin is associated with individual participants.

In the following study, a new dataset will be introduced to understand how MLMs can be used to improve the inferences made from a dataset. For example, given the model estimates, what are the range of likely effects possible in the population for a given behaviour? Such estimates are important when making policy decisions. When focusing on HF engineering outcomes,

such as the safety of transitions of control or driver state, the average is often of less importance, and what matters most for user safety, efficiency, and satisfaction is the worst-case extremes. The next section will also highlight what happens when things go wrong with MLMs. The example presented within Study 1 referred to models with only one continuous predictor variable. However, it is common in HF experiments to have more than one predictor variable, manipulated across two or more levels. The next section examines whether a MLM can be fitted with more than one random slope, and what happens when those slopes are unidentifiable.

3 Study 2: Gaze Entropy Metrics for Mental Workload (MWL) Estimation are Heterogenous During Hands-Off Level 2 Automation

3.1 Experimental description

This study focused on analysing how stationary gaze entropy (H_s) (Shiferaw et al. 2019) changed as a function of high MWL and other road traffic. H_s is a measure of the predictability of fixation locations, with a higher uncertainty (higher entropy) indicative of higher dispersion of fixations. H_s has been shown to reduce when drivers are under high MWL during manual (Schieber and Gilland 2008) and partially automated driving (Goodridge et al. 2024; Pillai et al. 2022). Participants were tasked with monitoring a hands-offs L2 driving system before transitions of control. The eye tracking data relates to the period of monitoring. The experiment had two independent variables manipulated over two levels: MWL (N-back/no N-back - Mehler et al. 2011) and lead vehicle presence (lead vehicle/no lead vehicle). The original manuscript (Goodridge et al. 2024) utilised a Bayesian distributional MLM, however the

current analysis focuses only on H_s using Frequentist MLMs[4]. The sample size used for the following sections is N = 38.

3.2 MLMs with dummy coded variables

To understand how H_s changed as a function of N-back and lead vehicle, an MLM can be fitted as follows:

$$H_{s_{ij}} \sim N(\mu_{ij}, \sigma_e) \tag{5}$$

$$\mu_{ij} = \left(\beta_0 + \beta_{0_j}\right) + \left(\beta_N N_i + \beta_{N_j} N_i\right) + (\beta_L L_i) + (\beta_{NL} NL_i)$$

$$\begin{bmatrix} \beta_{0_j} \\ \beta_{N_j} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, S\right)$$

$$S = \begin{pmatrix} \sigma^2_{\beta_{0_j}} & \rho_{N0}\sigma_{\beta_{N_j}}\sigma_{\beta_{0_j}} \\ \rho_{0N}\sigma_{\beta_{0_j}}\sigma_{\beta_{N_j}} & \sigma^2_{\beta_{N_j}} \end{pmatrix}$$

In Equation 5, as before, subscript i indexes observations and j indexes participants. The variable N_i indicates whether the N-back task was present during observation i, L_i indicates the presence of a lead vehicle, and NL_i represents the interaction between these two predictors. S denotes the variance-covariance matrix of the random effects, describing the variability in the participant-specific intercepts and slopes as well as their covariance. Random slopes (β_{N_j}) were included only for the N-back predictor in order to model potential between-participant variability in the effect of the N-back task on H_s. Random slopes for the lead vehicle predictor

[4] Frequentist models aim to find single point estimates for parameters that maximize the likelihood of the observed data, given the model. Bayesian models, in contrast, treat parameters as random variables with probability distributions that reflect uncertainty in their values (Nalborczyk et al. 2019). Bayesian approaches also use prior distributions to incorporate prior information about parameters.

and the interaction term were not included. The rationale for this modelling choice is discussed further in Section 3.5.

Both the N-back task and lead vehicle predictors are categorical variables with two levels: N-back/no N-back and lead vehicle/no lead vehicle. These predictors were represented using dummy coding, a common approach in regression models in which categorical variables are expressed as binary indicators taking values of 0 or 1. A reference level is specified (coded as 0), and the regression coefficients represent the change relative to this reference condition. In the present model, the N-back predictor was coded as $N_i \in \{0, 1\}$, where $N_i = 1$ indicates the presence of the N-back task. Similarly, the lead vehicle predictor was coded as $L_i \in \{0, 1\}$, where $L_i = 1$ indicates the presence of a lead vehicle during hands-off L2 driving. Under this coding scheme, the intercept parameter (β_0) represents the expected value of H_s in the reference condition where neither the N-back task nor the lead vehicle is present (N_i = 0, L_i = 0). The coefficient β_N represents the expected change in H_s associated with the presence of the N-back task relative to the no-N-back condition when no lead vehicle is present (Ben-Shachar, 2024). Similarly, β_L represents the expected change in H_s associated with the presence of a lead vehicle when the N-back task is absent. The interaction coefficient (β_{NL}) represents the additional change in H_s when both the N-back task and lead vehicle are present simultaneously, beyond the sum of their individual effects. In other words, it quantifies whether the effect of the N-back task on H_s differs depending on whether a lead vehicle is present. The summary of the fitted model is presented in Table 4.

Table 4: Summary of fixed and random effects for varying intercept-varying slope model for H_s

Fixed effects	Stationary gaze entropy (H_s)		
Predictors	*Estimates*	*CI*	*p*
β_0	.475	[-.433, -.518]	**<0.001**
β_N	-.142	[-.181, -.103]	**<0.001**
β_L	-.043	[-.061, -.024]	**<0.001**
β_{NL}	.019	[-.007, .045]	.158
Random Effects			
σ_e	.072		
$\sigma_{\beta_{0_j}}$	.122		
$\sigma_{\beta_{N_j}}$	.097		
$\rho_{0N}\sigma_{\beta_{0_j}}\sigma_{\beta_{N_j}}$	-0.34		
N	38		
Observations	744		

The model indicates a statistically significant effect for N-back induced MWL on H_s; for the average driver the presence of N-back reduces H_s by 14 percentage points (β_N = -.142) when the lead vehicle is not present. The confidence intervals suggest this could be as much as 18 percentage points, or as little as 10 percentage points. There was also a statistically significant effect of lead vehicle when the drivers were not completing N-back; H_s reduced by 4 percentage points on average. It should be noted that dummy coding is not the only way of handling categorical variables (see Ben-Shachar, 2024) although for "simple" research designs with definitive baseline conditions, dummy coding is perhaps the easiest way of interpreting model parameters.

3.3 Correlations

In Study 1, we alluded to the presence of correlation parameters being estimated between random effects. Many papers that use MLMs focus on the fixed effects, with a small minority interpreting the random slope and intercept parameters; papers rarely mention the correlation parameter. However, sometimes these parameters can be of inferential interest. The correlation parameter for the model specified in Equation 5 is estimated to be negative (e.g., $\rho_{0N}\sigma_{\beta_{0_j}}\sigma_{\beta_{N_j}}$ = -.34) (see Figure 4B). This suggests that individuals who have lower intercepts (i.e., lower H_s under no N-back and no lead vehicle trials) tend to have less steep β_{N_j} slopes (i.e., smaller reductions in H_s during N-back trials).

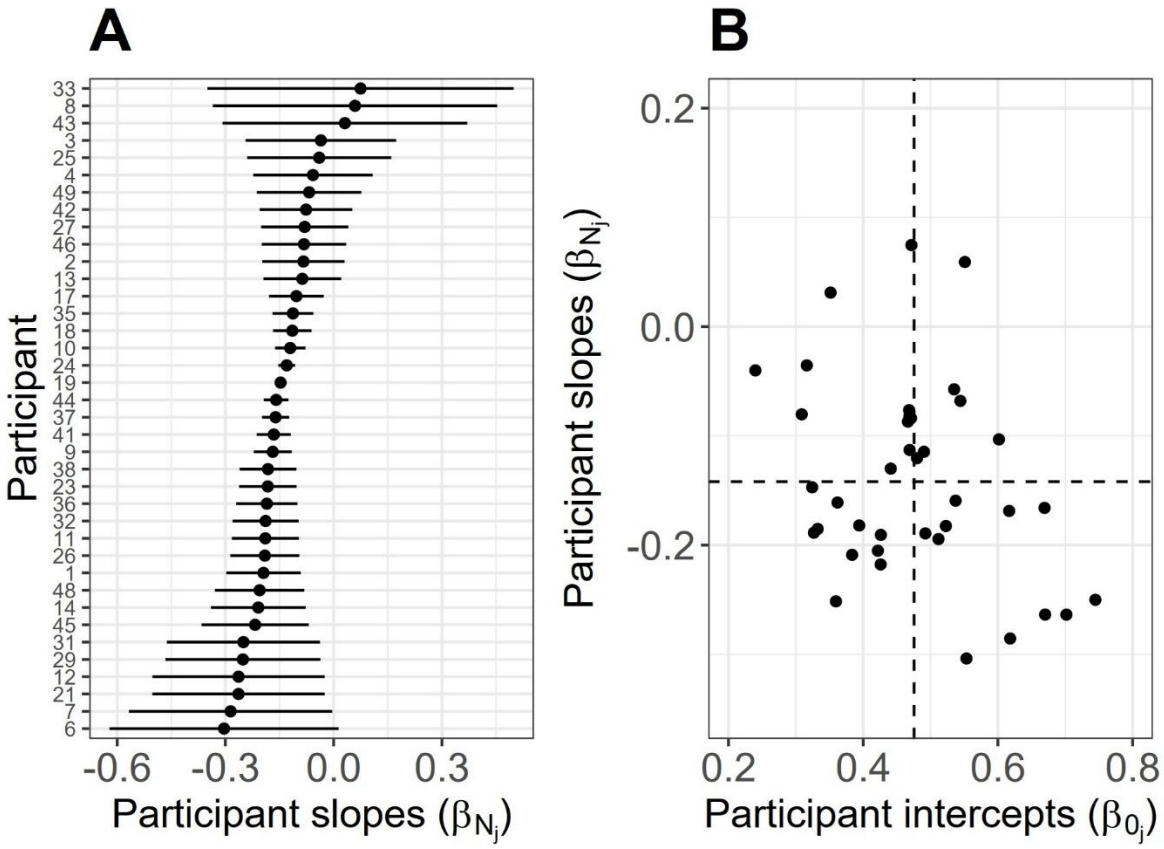

Figure 4: A) Participant random slopes (β_{N_j}) with bars displaying 95% confidence intervals B) Correlations between participant random intercepts (β_{0_j}) and random slopes (β_{N_j}). Each point presents an individual participant; dashed lines represent the grand intercept and slope.

Although correlation parameters are not typically the primary focus of many HF analyses, they can provide useful insight into individual differences. In the present model, the correlation between random intercepts and random slopes is of particular interest. A negative correlation indicates that participants with lower predicted H_s during baseline conditions (when both N-back and lead vehicle are absent, represented by smaller β_{0_j} values) tend to show smaller reductions in H_s in response to increased MWL. In this context, the random intercept represents individual differences in baseline H_s, while the random slope (β_{N_j}) represents individual differences in sensitivity of H_s to N-back-induced MWL. A negative intercept–slope correlation may therefore indicate a floor-like effect, whereby individuals who already exhibit low H_s under baseline conditions have limited capacity for further reductions when MWL increases. Such a pattern may also reflect measurement limitations. If H_s is already near a lower physiological or technological detection limit, further reductions may be difficult to detect using eye-tracking technology. Consequently, the magnitude of MWL-related changes in H_s may appear smaller for individuals with naturally low baseline H_s, potentially reducing the sensitivity of H_s as an indicator of MWL for these participants.

Another example could be the analysis of longitudinal data (Brown 2021). Consider a situation where a HF scientist conducts a longitudinal study on an individual's opinion on a piece of technology. In such a study, it might be of interest to know whether opinions of the technology at baseline are related to changes in opinions over time. The correlation between the participants specific intercepts (baseline opinions) and slopes (changes in opinion over time) can estimate this effect.

3.4 Heterogeneity intervals

The fixed effect model estimates suggest that, for the average driver, H_s reduces when MWL is high. One might conclude that H_s is useful candidate for future DMS for identifying whether

a driver is overloaded or not. However, basing this judgement on the average driver may not be the best strategy. Most people are not the average driver (e.g., only 16 participants had an N-back-induced reduction of H_s within ± 0.05 percentage points of the average), and thus an improved inference would be an estimate of the range of effects that are likely in the population, with a confidence of 95%. With this type of analysis, one can quantify the heterogeneity in the effect of high MWL on H_s.

Everything needed to calculate the heterogeneity in the effect of MWL is contained within the model specified in Equation 5. This model estimated a slope for each driver, assumed to be normally distributed in the population around a fixed effect (β_N) with a standard deviation ($\sigma_{\beta_{N_j}}$). With these parameters, it is possible to calculate *heterogeneity intervals* (Bolger et al. 2019). An excellent in-depth explanation of heterogeneity intervals has been provided previously (Bolger et al. 2019); we provide a summary here in the context of the current data. Whilst confidence intervals highlight variability in the *average effect* from one sample to another, heterogeneity intervals focus on the variability of the effect from one participant to another in the population. In other words, confidence intervals focus on the precision of the average effect; heterogeneity intervals focus on how participants might differ in the population. Calculating 95% heterogeneity intervals can be done via the following equation:

$$(\beta_N \pm 1.96 \times \sigma_{\beta_{N_j}}) \qquad 6$$

Applying this expression to the estimates in Table 4 produces a 95% heterogeneity interval of [-.332, .047]. This interval alone highlights that a one-size fits all interpretation of the model parameters would be insufficient for understanding whether H_s can be used as a measure of MWL. At one extreme, the model predicts that some drivers may have reductions in H_s that

are more than twice the size of the average effect. At the other extreme, the model predicts that drivers may have *increases* in H_s by up to 5 percentage points (see Figure 5).

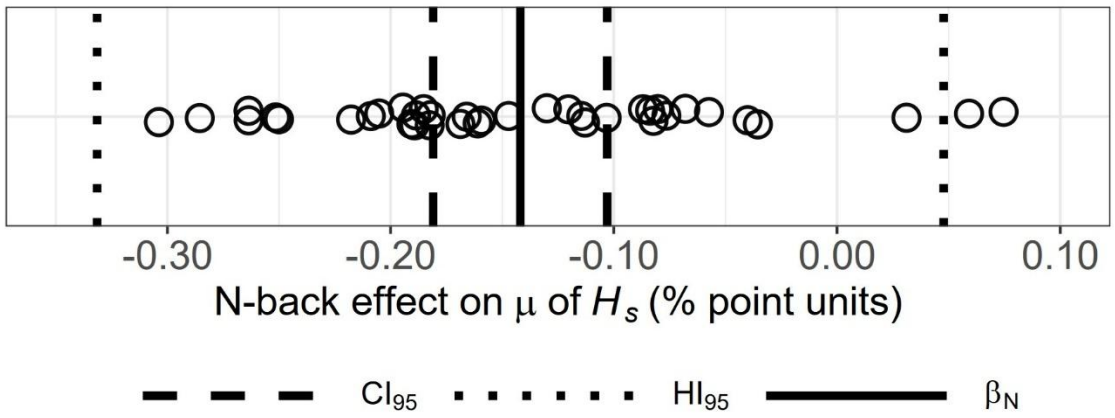

Figure 5: Strip plot displaying the range of effects of N-back on H_s. Open circles are fitted slopes for each participant. The solid line denotes the average decrease in H_s (fixed effect), the dashed lines denote the 95% confidence intervals, and the dotted lines denote the population heterogeneity of the effect of N-back.

The points plotted horizontally are the model estimates of the slopes for each driver; the black middle solid line represents the average effect of N-back on H_s; the dashed line represents the 95% confidence intervals; and the dotted lines represent the 95% heterogeneity intervals. With this information, it is important to understand whether the population-level heterogeneity is sufficient to be noteworthy. One heuristic is the range of heterogeneity intervals. A range that includes 0 indicates that heterogeneity is sufficient such that null effects and reversals of the main effects are possible in the population. As can be seen from Figure 5, our model did indeed estimate three participants to have such reversals. As a result, interpretation of the average effect would not be sufficient when understanding the population overall. If *increases* in H_s are predicted by the model, this could result in classifying participants as having optimal MWL despite them reporting high subjective MWL.

3.5 Maximal models and singular random effects

The random effects structure for the model specified in Equation 5 included random slopes for the MWL factor, but not for the lead vehicle factor (nor for their interaction). The decision process for specifying random effects is one of the hottest debates within the MLM literature. A “maximal” random effects structure includes random intercepts and random slopes for all within participant factors, as well as correlations between the random effect components (Barr et al., 2013). For Study 1, the maximal random effects structure was achieved in the varying intercept-varying slope model (Equation 4). This maximal random effects structure accounted for variability in each participant's random intercept and slope, as well as the correlation between these parameters. Adding additional experimental factors - and thus additional predictors into the model - means that maintaining a maximal random effects structure requires specifying not only a random slope for the new predictor, but also a random slope for the interaction between predictors, and a correlation term for each pair of random effect parameters.

It is worth taking a moment to understand the rationale for including additional random effects parameters when the number of predictors in the model increases. In Study 2, two independent variables were manipulated: MWL (N-back versus no N-back) and the presence of a lead vehicle (lead vehicle versus no lead vehicle). Therefore, participants may vary in the effect of the N-back, the lead vehicle, and interaction effects; and all these driver-specific differences may be correlated. There are two prominent schools of thought for selecting the random effects structure: maximal models or ‘sufficient’ models. Those in favour of maximal models include all potential random effects that may influence the results (Barr et al. 2013). In practice, when modelling the impact that lead vehicles have on driver gaze, a random slope allows the effect of lead vehicle to vary across participants; perhaps some drivers have reduced gaze dispersion due to focusing on the lead vehicle, while others show minimal changes. Leaving this out would

force the model to assume that lead vehicle presence affects everyone's gaze homogenously, with all of the real individual variation getting incorrectly folded into the model's error term. The knock-on effect is that results may be driven by a subset of drivers who are particularly sensitive to lead vehicles, rather than reflecting a consistent effect across drivers. This makes the model artificially confident in its estimate of the overall effect, making it more likely that a statistically significant effect is found when this is not the case (e.g., a Type I error). For Study 2, a maximal model would look like this:

$$H_{s_{ij}} \sim N(\mu_{ij}, \sigma_e) \tag{7}$$

$$\mu_{ij} = \left(\beta_0 + \beta_{0_j}\right) + \left(\beta_N N_i + \beta_{N_j} N_i\right) + \left(\beta_L L_i + \beta_{L_j} L_i\right) + \left(\beta_{NL} NL_i + \beta_{NL_j} NL_i\right)$$

$$\begin{bmatrix} \beta_{0_j} \\ \beta_{N_j} \\ \beta_{L_j} \\ \beta_{NL_j} \end{bmatrix} = MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, S\right)$$

$$S = \begin{pmatrix} \sigma^2_{\beta_{0_j}} & \rho_{0N}\sigma_{\beta_{0_j}}\sigma_{\beta_{N_j}} & \rho_{0L}\sigma_{\beta_{0_j}}\sigma_{\beta_{L_j}} & \rho_{0NL}\sigma_{\beta_{0_j}}\sigma_{\beta_{NL_j}} \\ \rho_{N0}\sigma_{\beta_{N_j}}\sigma_{\beta_{0_j}} & \sigma^2_{\beta_{N_j}} & \rho_{NL}\sigma_{\beta_{N_j}}\sigma_{\beta_{L_j}} & \rho_{NNL}\sigma_{\beta_{N_j}}\sigma_{\beta_{NL_j}} \\ \rho_{L0}\sigma_{\beta_{L_j}}\sigma_{\beta_{0_j}} & \rho_{LN}\sigma_{\beta_{L_j}}\sigma_{\beta_{N_j}} & \sigma^2_{\beta_{L_j}} & \rho_{LNL}\sigma_{\beta_{L_j}}\sigma_{\beta_{NL_j}} \\ \rho_{NL0}\sigma_{\beta_{NL_j}}\sigma_{\beta_{0_j}} & \rho_{NLN}\sigma_{\beta_{NL_j}}\sigma_{\beta_{N_j}} & \rho_{NLL}\sigma_{\beta_{NL_j}}\sigma_{\beta_{L_j}} & \sigma^2_{\beta_{NL_j}} \end{pmatrix}$$

The inclusion of an extra predictor variable results in two additional fixed effects parameters (i.e., the average effect of lead vehicle, β_L; and the interaction effect between N-back and lead vehicle, β_{NL}). These extra fixed effects increase the number of random effects needed to keep the model maximal; this is largely driven by the number of correlation parameters.

The second school of thought is that maximal models are frequently over-parameterised (Matuschek et al. 2017) and so just need to be ‘sufficient’. When random effects are included that account for very little real variance, model complexity is added without meaningful gain, and this comes with costs: maximal models often fail to converge, and when they do converge,

the random effect estimates can be implausible. Model convergence refers to a model that has successfully found the best estimates for the parameters it's trying to fit. During the fitting process, the model makes a series of adjustments to these estimates to minimise the error between the model's predictions and the data. Convergence means this process has stabilised and reached a point where further adjustments do not improve the model. If a model fails to converge, it suggests the fitting process got "stuck" and could not find reliable estimates. This might happen if the model is too complex, data are insufficient, or data are poorly scaled (i.e., predictors are on different scales and thus the variances are very different). Even if the model converges, complex random effects structures can produce "singularity" parameter estimates. This is where random effects are estimated as having a variability of exactly zero or correlations that are exactly +/-1. This can occur when specifying complex models with numerous random slope terms and/or when there is very little variability in the effect one is attempting to estimate. These singularity estimates are unreliable and uninterpretable, and over-parameterised models can reduce statistical power for testing fixed effects (Matuschek et al. 2017; Singmann and Kellen 2019).

Matuschek et al (2017) advocates for using likelihood ratio tests (LRTs) to compare models with and without random slopes, keeping only those that account for meaningful variance. Hence the random effects structure should be driven by the data, not by a blanket rule. The LRT asks whether a more complex model fits the data significantly better than a simpler one. If the answer is no, the random slope is dropped without meaningful cost to the model's accuracy. For example, a model would be fitted predicting H_s with lead vehicle as a fixed effect and a random slope for lead vehicle across participants, then comparing it against the same model without that random slope. If the LRT returns a non-significant result, it suggests participants are not varying enough to justify the added complexity, and the researcher should then proceed with the simpler model.

In practice, the sensible middle ground most researchers land on is a combination of the two: random effects are justified by theory and whether the current dataset support them. Any starting point for MLMs should be that your random effects structure reflects your understanding of how the data were generated, not just what the data happen to show. Before fitting a model, a researcher should be asking: is there a credible reason why different participants might respond differently to this predictor? In an experiment examining gaze, the answer is almost certainly yes. Research on driver gaze shows that eye movements are sensitive to a range of factors including NDRT engagement (Goodridge et al. 2024; Wilkie et al. 2019), driving mode (Deniel and Navarro 2023; Louw and Merat 2017), and driving experience (Babić et al. 2022; Mackenzie and Harris 2017) - all of which would plausibly moderate how strongly a lead vehicle impacts gaze behaviour. On those grounds alone, a random slope for lead vehicle is theoretically defensible and could be included to avoid Type I error inflation.

Once the model is fitted, it is worth inspecting the estimated variance of the random slope. If it is extremely small, or fitting the random effect has resulted in convergence issues, that is an empirical signal that the sample are not varying in their response, whatever the theory might suggest. In Study 2, the LRT between the model with and without the lead vehicle random effect reveals a non-significant effect [$\chi^2(3) = 1.04$, $p = .79$]. In this situation, retaining the lead vehicle random slope adds complexity without reflecting anything real in the data, and the parsimony argument begins to carry more weight. Hence the appropriate random effects structure is one that is both theoretically motivated and empirically supported, rather than one derived mechanically from either principle alone (Matuschek et al. 2017)[5].

[5] In addition to LRTs, model selection can also be conducted using information criteria such as the Akaike Information Criterion (AIC) (Akaike 1998) or Bayesian Information Criterion (BIC) (Schwarz 1978), which

4 Discussion

This manuscript provides a tutorial on how to use MLMs on HF datasets. Study 1 illustrated how MLMs are effectively an extension of ordinary linear regression, such that individual participants have their own intercept and slope which allows a researcher to estimate the variability in an effect of interest. Study 2 built on this by demonstrating how explicitly modelling human response heterogeneity provides additional insight on the size and variation of experimental effects within our experiments. These examples demonstrate how MLMs provide a flexible framework for analysing complex, hierarchical, and heterogeneous data structures commonly encountered in HF research.

The application of MLMs has the potential to be far reaching if HF scientists have the resources, motivations, and skillsets to implement them. For example, Freed et al (2021) pointed out that although some naturalistic driving studies use MLMs to account for nested data structures (Ellison et al. 2015; Klauer et al. 2014), they often do not characterise the variability as an important descriptive or outcome variable. This represents a lost opportunity for theory development or system re-designs to account for these idiosyncrasies. One study found that 29.9% total truck driver route time variation was attributed to the individual truck drivers themselves (Keil et al. 2024). Acknowledging the significant influence of individual behaviour on overall performance can enable managers to make more informed and effective decisions, ultimately improving both process efficiency and driver well-being. For instance, performance analysis could be used to identify sub-groups of drivers who may benefit from tailored training and targeted support initiatives. Within the context of DMS, if we are to identify useful metrics of MWL, not only do they need to be valid (i.e., correlate with subjective

balance model fit and complexity. The discussion of these is beyond the scope of this manuscript, however these have been covered in the literature (Matuschek et al. 2017).

MWL assessments) but they also need to be reliable (i.e., change as a function of MWL in similar ways across a population). Therefore, between participant variability is fundamental in generating a better understanding of human responses. Information about this heterogeneity is useful if an analysis technique can incorporate it.

To illustrate how MLMs can improve the statistical inferences available to HF scientists, a comparison between the outputs of a RM ANOVA and a MLM is provided akin to that highlighted by Bolger et al (2019). The following will use a research question that was proposed in Study 2. A HF scientist wants to understand how MWL influences H_s. Using a conventional analytic framework:

"A 2 x2 Repeated Measures ANOVA was conducted to investigate the effect of lead vehicle and N-back on H_s. There was a significant main effect of N-back [F (1, 36) = 50.372, p < 0.001]. This suggests that the gaze of drivers was significantly less dispersed when completing N-back (M = 0.313, SD = 0.142) versus when monitoring a hands-off Level 2 system (M = 0.443, SD = 0.143) with a very large effect size (η_p^2 = 0.58). This significant effect suggests that the distribution of driver's gaze is reduced when they are under high MWL. As such, H_s is a variable that shows great promise in estimating high MWL in drivers using hands-off Level 2 driving systems".

This analysis is instantly recognisable within the HF literature; however, it does not provide the full picture. Using an MLM identifies the mean difference in H_s, but builds on it to provide a richer inference:

"A multilevel model was fitted to investigate the effect of lead vehicle and N-back on H_s. The model revealed a significant main effect of MWL on H_s (β_N = -0.141, CI_{95} = [-0.180, -0.102], p < .001). This suggests that a typical driver's gaze is significantly less dispersed when completing N-back (M = 0.313, SD = 0.142) versus when monitoring a hands-off Level

2 system (M = 0.443, SD = 0.143). For the average driver this effect is expected to be a reduction of H_s of around 14 percentage points. However, the multilevel model revealed that this effect is not expected to be homogenous across the population. The random effects within the model imply that some drivers can be expected to have reductions in H_s of 33 percentage points, whilst other drivers are expected to have no change in H_s or even slight reversals of the average effect (HI_{95} = [-0.331, 0.047]). Whilst H_s seems suitable for estimating MWL for the typical driver, there are many drivers in the population who may experience high MWL but would not be detected using this metric".

Some HF fields are already reaping the benefits of exploring heterogeneity in their analyses. In logistics research, analyses of forklift batch execution times (Loske and Klumpp 2022; Matusiak et al. 2017) and truck pick up times (Keil et al. 2024) have found that individuals in the system can account for considerable amounts of variance in an outcome variable. This reinforces the principles that logistics system outcomes cannot be planned and evaluated comprehensively without considering workforce heterogeneity.

In aviation research, MLMs are rarely used despite experiments producing data structures well suited to them (e.g., repeated measures and small samples) (Lorenz et al. 2024). One exception used an MLM to combine data from two flight-simulation experiments (Zaal et al. 2021). Although each experiment had a small sample, the pooled analysis replicated the original RM-ANOVA results while enabling meaningful comparisons across studies. This illustrates another advantage of MLMs: the ability to pool datasets to increase statistical power. While cross-study pooling is common in meta-analyses (Hox 2002; Raudenbush 2002), it typically relies on summary statistics rather than raw data. Recent work suggests MLMs could be used to pool raw datasets directly (Goodridge et al. 2025), although limited data sharing in HF research remains a barrier (Ebel et al. 2024)

MLMs can also improve research practises more broadly. Replication science refers to the systematic investigation of previous work to assess the validity and reliability of scientific findings (Nosek and Errington 2020). In fields such as Human-Technology Interaction (HTI), not only are replication studies undervalued (Hornbæk et al. 2014; Wilson et al. 2011), but when they are conducted, the original results often do not replicate (Leichtmann and Nitsch 2021; Ullman et al. 2021). Replication failures are a concern (Shrout and Rodgers 2018) and there is evidence to suggest that replication studies from heterogeneous populations are less likely to detect true effects (Bolger et al. 2019; Maxwell et al. 2017). Hence an important implication of heterogenous effects is that larger sample sizes will be needed to maintain statistical power. However, because MLMs estimate the size and heterogeneity of a particular effect, they can be used to assess whether failures in replication are due to true null effects in the population or due to heterogeneity. It was recently suggested that a reason why MWL may have inconsistent impact on the quality of transitions from automated driving could be due to the population level effect size being small and effects being heterogenous in the population (Goodridge et al. 2026). There is also heterogeneity associated with methodologies used to investigate the impacts of NDRTs during automated driving (Wijayaratna et al. 2019). In short, even if a HF scientist wishes to focus purely on the average effects of an experimental manipulation, this approach should be justified using a MLM to highlight that there is minimal heterogeneity, or that sample heterogeneity can be safely ignored.

Despite the advantages of MLMs, it is important to recognise that statistical models remain simplified representations of complex real-world processes. Random effects improve analyses by better approximating the underlying process for how data are generated. However, even MLMs only capture a subset of the variability present in most experiments. Many sources of variation - differences in cognitive ability, experience, or demographic background - often remain unmodelled. Similarly, while MLMs allow inference beyond the sampled participants,

this generalisation is limited by the populations represented in the data. Consequently, the validity of statistical inference depends not only on the model itself but also on the researcher's intended scope of generalisation. Aligning modelling choices with these inferential goals is essential to avoid misleading conclusions (Cornfield and Tukey 1956; Yarkoni 2022).

In conclusion, when more than one observation occurs for the same individual, or when different individuals have differing numbers of observations, an MLM allows these elements to be accounted for where a traditional single level model (i.e., a linear regression) might mislead (McElreath 2018). Furthermore, MLMs allow us to answer research questions regarding variations in the effects under study. Understanding variation is key in HF research, given that the negative safety implications occur within the tails of a response distribution. MLMs are a big help, as they explicitly model this variation. Finally, they help avoid the averaging that is necessitated by the traditional single level models. Averaging removes variance and thus produces false confidence in estimates. MLMs maintain the variation contained within pre-averaged data, whilst using the average to make predictions. McElreath (2018) proposed that MLMs should be the default form of regression, and deviations from this default should be justified. We agree with this notion and have provided a detailed tutorial on how to use these models on HF data. We hope this is the beginning of a change toward wider adoption of these research methods, fitting MLMs, and fully utilising the parameters to produce better inferences from our data.

5 References

Akaike, Hirotogu. 1998. 'Information Theory and an Extension of the Maximum Likelihood Principle'. In *Selected Papers of Hirotugu Akaike*, edited by Emanuel Parzen, Kunio Tanabe, and Genshiro Kitagawa. Springer Series in Statistics. Springer New York. https://doi.org/10.1007/978-1-4612-1694-0_15.

Baayen, R. Harald. 2004. ‘Statistics in Psycholinguistics: A Critique of Some Current Gold Standards’. *Mental Lexicon Working Papers* 1 (1): 1–47.

Babić, Darko, Ivana Prelčec, Dario Babić, Vanna Boroša, and Mario Fiolić. 2022. ‘The Impact of Mobile Phone Use on Young Drivers’ Driving Behaviour and Visual Scanning of the Environment’. *Promet-Traffic&Transportation* 34 (3): 431–41.

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. ‘Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal’. *Journal of Memory and Language* 68 (3): 255–78.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. ‘Fitting Linear Mixed-Effects Models Using Lme4’. *Journal of Statistical Software* 67: 1–48.

Boer, Erwin R. 2016. ‘Satisficing Curve Negotiation: Explaining Drivers’ Situated Lateral Position Variability’. *IFAC-PapersOnLine* 49 (19): 183–88.

Boisgontier, Matthieu P., and Boris Cheval. 2016. ‘The Anova to Mixed Model Transition’. *Neuroscience & Biobehavioral Reviews* 68: 1004–5.

Bolger, Niall, Katherine S. Zee, Maya Rossignac-Milon, and Ran R. Hassin. 2019. ‘Causal Processes in Psychology Are Heterogeneous.’ *Journal of Experimental Psychology: General* 148 (4): 601.

Brown, Violet A. 2021. ‘An Introduction to Linear Mixed-Effects Modeling in R’. *Advances in Methods and Practices in Psychological Science* 4 (1): 2515245920960351. https://doi.org/10.1177/2515245920960351.

Bürkner, Paul-Christian. 2017. ‘Brms: An R Package for Bayesian Multilevel Models Using Stan’. *Journal of Statistical Software* 80: 1–28.

Chen, Jessie Y. C., Stephanie A. Quinn, Julia L. Wright, and Michael J. Barnes. 2013. 'Effects of Individual Differences on Human-Agent Teaming for Multi-Robot Control'. In *Engineering Psychology and Cognitive Ergonomics. Understanding Human Cognition*, edited by Don Harris, vol. 8019, edited by David Hutchison, Takeo Kanade, Josef Kittler, et al. Lecture Notes in Computer Science. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-39360-0_30.

Clegg, Chris W. 2000. 'Sociotechnical Principles for System Design'. *Applied Ergonomics* 31 (5): 463–77.

Cornfield, Jerome, and John W. Tukey. 1956. 'Average Values of Mean Squares in Factorials'. *The Annals of Mathematical Statistics*, 907–49.

Davidian, Marie, and David M. Giltinan. 2003. 'Nonlinear Models for Repeated Measurement Data: An Overview and Update'. *Journal of Agricultural, Biological, and Environmental Statistics* 8 (4): 387.

Davies, G. Matt, and Alan Gray. 2015. 'Don't Let Spurious Accusations of Pseudoreplication Limit Our Ability to Learn from Natural Experiments (and Other Messy Kinds of Ecological Monitoring)'. *Ecology and Evolution* 5 (22): 5295–304. https://doi.org/10.1002/ece3.1782.

Deniel, Jonathan, and Jordan Navarro. 2023. 'Gaze Behaviours Engaged While Taking over Automated Driving: A Systematic Literature Review'. *Theoretical Issues in Ergonomics Science* 24 (1): 54–87. https://doi.org/10.1080/1463922X.2022.2036861.

Dunning, Richard E., Baruch Fischhoff, and Alex L. Davis. 2024. 'When Do Humans Heed AI Agents' Advice? When Should They?' *Human Factors: The Journal of the Human*

Factors and Ergonomics Society 66 (7): 1914–27. https://doi.org/10.1177/00187208231190459.

Ebel, Patrick, Pavlo Bazilinskyy, Mark Colley, et al. 2024. 'Changing Lanes Toward Open Science: Openness and Transparency in Automotive User Research'. *Proceedings of the 16th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, September 22, 94–105. https://doi.org/10.1145/3640792.3675730.

Ellison, Adrian B., Stephen P. Greaves, and Michiel CJ Bliemer. 2015. 'Driver Behaviour Profiles for Road Safety Analysis'. *Accident Analysis & Prevention* 76: 118–32.

Freed, Sara A., Lesley A. Ross, Alyssa A. Gamaldo, and Despina Stavrinos. 2021. 'Use of Multilevel Modeling to Examine Variability of Distracted Driving Behavior in Naturalistic Driving Studies'. *Accident Analysis & Prevention* 152: 105986.

Gajera, Hardik, Srinivas S. Pulugurtha, and Sonu Mathew. 2022. *Influence of Level 1 and Level 2 Automated Vehicles on Fatal Crashes and Fatal Crash Occurrence*. https://scholarworks.sjsu.edu/mti_publications/402/.

Godthelp, Hans, Paul Milgram, and Gerard J. Blaauw. 1984. 'The Development of a Time-Related Measure to Describe Driving Strategy'. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 26 (3): 257–68. https://doi.org/10.1177/001872088402600302.

Goodridge, Courtney M., Rafael C. Goncalves, Ali Arabian, et al. 2024. 'Gaze Entropy Metrics for Mental Workload Estimation Are Heterogenous during Hands-off Level 2 Automation'. *Accident Analysis & Prevention* 202: 107560.

Goodridge, Courtney M., Rafael C. Gonçalves, Ali Arabian, et al. 2026. 'The Impact of N-Back-Induced Mental Workload and Time Budget on Takeover Performance'. *Accident Analysis & Prevention* 225: 108327.

Goodridge, Courtney M., Rafael C. Gonçalves, Amélie Reher, Jonny Kuo, Michael G. Lenné, and Natasha Merat. 2025. 'Assessing Data Imbalance Correction Methods and Gaze Entropy for Collision Prediction'. *PloS One* 20 (11): e0336777.

Grassmann, Mariel, Elke Vlemincx, Andreas von Leupoldt, and Omer Van den Bergh. 2017. 'Individual Differences in Cardiorespiratory Measures of Mental Workload: An Investigation of Negative Affectivity and Cognitive Avoidant Coping in Pilot Candidates'. *Applied Ergonomics* 59: 274–82.

Grech, Michelle R., Andrew Neal, Gillian Yeo, Michael Humphreys, and Simon Smith. 2009. 'An Examination of the Relationship between Workload and Fatigue within and across Consecutive Days of Work: Is the Relationship Static or Dynamic?' *Journal of Occupational Health Psychology* 14 (3): 231.

Harrison, Xavier A., Lynda Donaldson, Maria Eugenia Correa-Cano, et al. 2018. 'A Brief Introduction to Mixed Effects Modelling and Multi-Model Inference in Ecology'. *PeerJ* 6: e4794.

Haverkamp, Nicolas, and André Beauducel. 2017. 'Violation of the Sphericity Assumption and Its Effect on Type-I Error Rates in Repeated Measures ANOVA and Multi-Level Linear Models (MLM)'. *Frontiers in Psychology* 8: 1841.

Hornbæk, Kasper, Søren S. Sander, Javier Andrés Bargas-Avila, and Jakob Grue Simonsen. 2014. 'Is Once Enough?: On the Extent and Content of Replications in Human-

Computer Interaction'. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, April 26, 3523–32. https://doi.org/10.1145/2556288.2557004.

Hox, Joop J. 2002. 'Multilevel Multivariate and Structural Equation Models: Some Missing Links'. *Social Science Methodology in the New Millennium. Opladen, Germany: Leske+ Budrich*. http://www.joophox.net/papers/p090103.pdf.

Jipp, Meike. 2014. 'Levels of Automation: Effects of Individual Differences on Wheelchair Control Performance and User Acceptance'. *Theoretical Issues in Ergonomics Science* 15 (5): 479–504. https://doi.org/10.1080/1463922X.2013.815829.

Kamaraj, Amudha V., Joonbum Lee, Joshua E. Domeyer, Shu-Yuan Liu, and John D. Lee. 2024. 'Comparing Subjective Similarity of Automated Driving Styles to Objective Distance-Based Similarity'. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 66 (5): 1545–63. https://doi.org/10.1177/00187208221142126.

Keil, Maria, Dominic Loske, Tiziana Modica, and Matthias Klumpp. 2024. 'Human Factors on the Road: Truck Drivers' Heterogeneity in Distribution'. *IFAC-PapersOnLine* 58 (19): 964–69.

Kim, Jin Yong, Corey Lester, and X. Jessie Yang. 2025. 'Beyond Binary Decisions: Evaluating the Effects of AI Error Type on Trust and Performance in AI-Assisted Tasks'. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 67 (10): 1062–83. https://doi.org/10.1177/00187208251326795.

Klauer, Sheila G., Feng Guo, Bruce G. Simons-Morton, Marie Claude Ouimet, Suzanne E. Lee, and Thomas A. Dingus. 2014. 'Distracted Driving and Risk of Road Crashes among Novice and Experienced Drivers'. *New England Journal of Medicine* 370 (1): 54–59.

Kliegl, Reinhold, Ping Wei, Michael Dambacher, Ming Yan, and Xiaolin Zhou. 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.

Körber, Moritz, and Klaus Bengler. 2014. 'Potential Individual Differences Regarding Automation Effects in Automated Driving'. *Proceedings of the XV International Conference on Human Computer Interaction*, September 10, 1–7. https://doi.org/10.1145/2662253.2662275.

Lansdown, Terry C. 2002. 'Individual Differences during Driver Secondary Task Performance: Verbal Protocol and Visual Allocation Findings'. *Accident Analysis & Prevention* 34 (5): 655–62.

Leichtmann, Benedikt, and Verena Nitsch. 2021. 'Is the Social Desirability Effect in Human–Robot Interaction Overestimated? A Conceptual Replication Study Indicates Less Robust Effects'. *International Journal of Social Robotics* 13 (5): 1013–31. https://doi.org/10.1007/s12369-020-00688-z.

Lin, Jinchao. 2017. *The Impact of Automation and Stress on Human Performance in UAV Operation*. https://stars.library.ucf.edu/etd/5710/.

Lo, Steson, and Sally Andrews. 2015. 'To Transform or Not to Transform: Using Generalized Linear Mixed Models to Analyse Reaction Time Data'. *Frontiers in Psychology* 6: 1171.

Lorenz, Bernd, Catherine Chalon-Morgan, Ivan De Visscher, and Thomas Feuerle. 2024. 'Application of Linear Mixed-Effect Modeling for the Analysis of Human-in-the-

Loop Simulation Experiments'. *Journal of Air Transportation* 32 (2): 71–83. https://doi.org/10.2514/1.D0394.

Loske, Dominic, and Matthias Klumpp. 2022. 'Quantifying Heterogeneity in Human Behavior: An Empirical Analysis of Forklift Operations through Multilevel Modeling'. *Logistics Research* 15 (1): 1–17.

Louw, Tyron, Jonny Kuo, Richard Romano, Vishnu Radhakrishnan, Michael G. Lenné, and Natasha Merat. 2019. 'Engaging in NDRTs Affects Drivers' Responses and Glance Patterns after Silent Automation Failures'. *Transportation Research Part F: Traffic Psychology and Behaviour* 62: 870–82.

Louw, Tyron, and Natasha Merat. 2017. 'Are You in the Loop? Using Gaze Dispersion to Understand Driver Visual Attention during Vehicle Automation'. *Transportation Research Part C: Emerging Technologies* 76: 35–50.

Mackenzie, Andrew K., and Julie M. Harris. 2017. 'A Link between Attentional Function, Effective Eye Movements, and Driving Ability.' *Journal of Experimental Psychology: Human Perception and Performance* 43 (2): 381.

Malik, Jeehan, Nam-Yoon Kim, Morgan D. N. Parr, Joseph K. Kearney, Jodie M. Plumert, and Kyle Rector. 2024. 'Do Simulated Augmented Reality Overlays Influence Street-Crossing Decisions for Non-Mobility-Impaired Older and Younger Adult Pedestrians?' *Human Factors: The Journal of the Human Factors and Ergonomics Society* 66 (5): 1520–30. https://doi.org/10.1177/00187208231151280.

Malik, Jeehan, Elizabeth O'Neal, Megan Noonan, et al. 2025. 'Do Augmented Reality Cues Aid Pedestrians in Crossing Multiple Lanes of Traffic? A Virtual Reality Study'.

Human Factors: The Journal of the Human Factors and Ergonomics Society 67 (8): 823–35. https://doi.org/10.1177/00187208251320907.

Matthews, Gerald, Peter A. Hancock, Jinchao Lin, et al. 2021. 'Evolution and Revolution: Personality Research for the Coming World of Robots, Artificial Intelligence, and Autonomous Systems'. *Personality and Individual Differences* 169: 109969.

Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen, and Douglas Bates. 2017. 'Balancing Type I Error and Power in Linear Mixed Models'. *Journal of Memory and Language* 94: 305–15.

Matusiak, Marek, René De Koster, and Jari Saarinen. 2017. 'Utilizing Individual Picker Skills to Improve Order Batching in a Warehouse'. *European Journal of Operational Research* 263 (3): 888–99.

Maxwell, Scott E., Harold D. Delaney, and Ken Kelley. 2017. *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Routledge. https://api.taylorfrancis.com/content/books/mono/download?identifierName=doi&identifierValue=10.4324/9781315642956&type=googlepdf.

McElreath, Richard. 2018. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC. https://api.taylorfrancis.com/content/books/mono/download?identifierName=doi&identifierValue=10.1201/9781315372495&type=googlepdf.

Mehler, Bruce, Bryan Reimer, and Jeffery Dusek. 2011. 'MIT AgeLab Delayed Digit Recall Task (n-Back)'. *MIT Center for Transportation & Logistics Research Paper*, no. 2011/018. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6206338.

Mole, Callum D., Otto Lappi, Oscar Giles, Gustav Markkula, Franck Mars, and Richard M. Wilkie. 2019. 'Getting Back Into the Loop: The Perceptual-Motor Determinants of Successful Transitions out of Automated Driving'. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 61 (7): 1037–65. https://doi.org/10.1177/0018720819829594.

Muradoglu, Melis, Joseph R. Cimpian, and Andrei Cimpian. 2023. 'Mixed-Effects Models for Cognitive Development Researchers'. *Journal of Cognition and Development* 24 (3): 307–40. https://doi.org/10.1080/15248372.2023.2176856.

Nalborczyk, Ladislas, Cédric Batailler, Hélène Lœvenbruck, Anne Vilain, and Paul-Christian Bürkner. 2019. 'An Introduction to Bayesian Multilevel Models Using Brms: A Case Study of Gender Effects on Vowel Variability in Standard Indonesian'. *Journal of Speech, Language, and Hearing Research* 62 (5): 1225–42. https://doi.org/10.1044/2018_JSLHR-S-18-0006.

Nosek, Brian A., and Timothy M. Errington. 2020. 'What Is Replication?' *PLoS Biology* 18 (3): e3000691.

Park, Jungkyu, Ramsey Cardwell, and Hsiu-Ting Yu. 2020. 'Specifying the Random Effect Structure in Linear Mixed Effect Models for Analyzing Psycholinguistic Data'. *Methodology* 16 (2): 92–111.

Pillai, Prarthana, Balakumar Balasingam, Yong Hoon Kim, Chris Lee, and Francesco Biondi. 2022. 'Eye-Gaze Metrics for Cognitive Load Detection on a Driving Simulator'. *Ieee/Asme Transactions On Mechatronics* 27 (4): 2134–41.

Raudenbush, Stephen W. 2002. 'Hierarchical Linear Models: Applications and Data Analysis Methods'. *Advanced Quantitative Techniques in the Social Sciences Series/SAGE*.

https://books.google.com/books?hl=en&lr=&id=uyCV0CNGDLQC&oi=fnd&pg=PR17&dq=Raudenbush+%26+Bryk,+2002&ots=qD8FXnV1SG&sig=hC3cMfNKyb6BEfLmfkTb1G5_NVg.

Royle, J. Andrew. 2013. 'Review of: Mixed Effects Models and Extensions in Ecology with R'. arXiv:1305.6995. Preprint, arXiv, May 30. https://doi.org/10.48550/arXiv.1305.6995.

Schieber, Frank, and Jess Gilland. 2008. 'Visual Entropy Metric Reveals Differences in Drivers' Eye Gaze Complexity across Variations in Age and Subsidiary Task Load'. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52 (23): 1883–87. https://doi.org/10.1177/154193120805202311.

Schwarz, Gideon. 1978. 'Estimating the Dimension of a Model'. *The Annals of Statistics*, 461–64.

Shiferaw, Brook, Luke Downey, and David Crewther. 2019. 'A Review of Gaze Entropy as a Measure of Visual Scanning Efficiency'. *Neuroscience & Biobehavioral Reviews* 96: 353–66.

Shin, Juh Hyun. 2009. 'Application of Repeated-Measures Analysis of Variance and Hierarchical Linear Model in Nursing Research'. *Nursing Research* 58 (3): 211–17.

Shrout, Patrick E., and Joseph L. Rodgers. 2018. 'Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis'. *Annual Review of Psychology* 69 (1): 487–510. https://doi.org/10.1146/annurev-psych-122216-011845.

Singmann, Henrik, and David Kellen. 2019. 'An Introduction to Mixed Models for Experimental Psychology'. In *New Methods in Cognitive Psychology*. Routledge.

https://www.taylorfrancis.com/chapters/edit/10.4324/9780429318405-2/introduction-mixed-models-experimental-psychology-henrik-singmann-david-kellen.

Stawski, Robert S. 2013. 'Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling (2nd Edition)'. *Structural Equation Modeling: A Multidisciplinary Journal* 20 (3): 541–50. https://doi.org/10.1080/10705511.2013.797841.

Ullman, Daniel, Salomi Aladia, and Bertram F. Malle. 2021. 'Challenges and Opportunities for Replication Science in HRI: A Case Study in Human-Robot Trust'. *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, March 8, 110–18. https://doi.org/10.1145/3434073.3444652.

Wiebel-Herboth, Christiane B., Matti Krüger, and Patricia Wollstadt. 2021. 'Measuring Inter- and Intra-Individual Differences in Visual Scan Patterns in a Driving Simulator Experiment Using Active Information Storage'. *PLoS One* 16 (3): e0248166.

Wijayaratna, Kasun P., Mitchell L. Cunningham, Michael A. Regan, Sisi Jian, Sai Chand, and Vinayak V. Dixit. 2019. 'Mobile Phone Conversation Distraction: Understanding Differences in Impact between Simulator and Naturalistic Driving Studies'. *Accident Analysis & Prevention* 129: 108–18.

Wilkie, Richard, Callum Mole, Oscar Giles, Natasha Merat, Richard Romano, and Gustav Makkula. 2019. 'Cognitive Load during Automation Affects Gaze Behaviours and Transitions to Manual Steering Control'. *Driving Assessment Conference* 10 (2019). https://pubs.lib.uiowa.edu/driving/article/id/28372/.

Wilson, Max L., Wendy Mackay, Ed Chi, Michael Bernstein, Dan Russell, and Harold Thimbleby. 2011. 'RepliCHI - CHI Should Be Replicating and Validating Results

More: Discuss'. *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, May 7, 463–66. https://doi.org/10.1145/1979742.1979491.

Winter, Bodo. 2019. *Statistics for Linguists: An Introduction Using R*. Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9781315165547/statistics-linguists-introduction-using-bodo-winter.

Yarkoni, Tal. 2022. 'The Generalizability Crisis'. *Behavioral and Brain Sciences* 45: e1.

Zaal, Peter, Daan M. Pool, and Max Mulder. 2021. 'Linear Mixed-Effects Models for Human-in-the-Loop Tracking Experiment Data'. Paper presented at AIAA Scitech 2021 Forum. *AIAA Scitech 2021 Forum*, January 11. https://doi.org/10.2514/6.2021-1014.

Zgonnikov, Arkady, David Abbink, and Gustav Markkula. 2024. 'Should I Stay or Should I Go? Cognitive Modeling of Left-Turn Gap Acceptance Decisions in Human Drivers'. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 66 (5): 1399–413. https://doi.org/10.1177/00187208221144561.