Linear Mixed-Effects Models for Human-in-the-Loop Tracking Experiment Data

Peter M. T. Zaal*
Metis Technology Solutions, Inc.
NASA Ames Research Center
Moffett Field, CA, USA

Daan M. Pool†
Delft University of Technology
Delft, The Netherlands

Max Mulder‡
Delft University of Technology
Delft, The Netherlands

Linear mixed-effects models provide several benefits over more traditional statistical inference tests and are particularly useful for most human-in-the-loop tracking experiment data. However, surprisingly, mixed models are virtually not used for the analysis of tracking experiment data. This paper uses linear mixed-effects models to analyze combined tracking data from two previous human-in-the-loop roll tracking experiments that compared control behavior metrics collected in both a research aircraft and a motion-base simulator. In the experiments, pilots' behavior under 10 different motion configurations with varying motion filter gains and break frequencies was evaluated and compared to that in the real aircraft. The linear mixed-effects model analysis on the combined dataset confirmed the main statistical outcomes of the individual experiments. The main benefits of mixed models for this type of data were demonstrated by successfully combining data from two experiments that used different experimental conditions and of which one had an additional apparatus and the other a missing participant. Finally, the mixed-model analysis was able to explicitly test for scientifically relevant statistical differences in the dependent measures between the aircraft and simulator, as well as between both experiments.

Nomenclature

e	=	tracking error signal, deg
e_s	=	simulator tracking error signal, deg
F	=	ANOVA F-value
f_d	=	disturbance forcing function, V
f_t	=	target forcing function, deg
H_c	=	controlled element dynamics
H_{mf}	=	motion filter dynamics
H_{nm}	=	neuromuscular actuation dynamics
H_{p_m}	=	pilot motion response
H_{p_v}	=	pilot visual response
H_{s_m}	=	simulator motion cueing system dynamics
H_{s_v}	=	simulator visual cueing system dynamics
K_m	=	pilot motion gain, V/IPUT
K_{mf}	=	motion filter gain
K_s	=	stick input gain
K_v	=	pilot visual gain, V/deg
n	=	pilot remnant signal, V
p	=	statistical p-value
s	=	Laplace operator, rad/s
T_L	=	pilot visual lead time constant, s
u	=	pilot control signal, V
$\boldsymbol{u}$	=	vector of random effects
u_c	=	scaled pilot control signal, V
u_m	=	pilot motion control signal, V
u_v	=	pilot visual control signal, V
X	=	fixed effect design matrix
$\boldsymbol{y}$	=	observation vector
Z	=	random effect design matrix

Symbols

$\boldsymbol{\beta}$	=	vector of fixed effects
δ_c	=	control input, deg
$\boldsymbol{\epsilon}$	=	vector of random modeling errors
ζ_{nm}	=	neuromuscular damping ratio
λ	=	Box-Cox transformation factor
σ_e^2	=	error variance, deg^2
σ_u^2	=	control input variance, V^2
$\sigma_{u_m}^2/\sigma_{u_v}^2$	=	pilot motion/visual variance fraction

*Principal Aerospace Engineer, SimLabs, NASA Ames Research Center, Moffett Field, CA, 94035; peter.m.t.zaal@nasa.gov. Associate Fellow AIAA.

†Assistant Professor, Control & Simulation Section, Faculty of Aerospace Engineering, P.O. Box 5058, 2600GB Delft, The Netherlands; d.m.pool@tudelft.nl. Senior Member AIAA.

‡Professor, Control & Simulation Section, Faculty of Aerospace Engineering, P.O. Box 5058, 2600GB Delft, The Netherlands; m.mulder@tudelft.nl. Associate Fellow AIAA.

τ_m	=	pilot motion time delay, s	$\varphi_{m,t}$	=	target phase margin, deg
τ_v	=	pilot visual time delay, s	χ^2	=	likelihood ratio test statistic
ϕ	=	roll angle, deg	$\omega_{c,d}$	=	disturbance crossover frequency, rad/s
$\ddot{\phi}$	=	roll acceleration, deg/s^2	$\omega_{c,t}$	=	target crossover frequency, rad/s
$\ddot{\phi}_{mf}$	=	filtered roll acceleration, deg/s^2	ω_{mf}	=	motion filter break frequency, rad/s
$\ddot{\phi}_s$	=	simulator roll acceleration, deg/s^2	ω_{nm}	=	neuromuscular frequency, rad/s
$\varphi_{m,d}$	=	disturbance phase margin, deg			

I. Introduction

Statistical inference is an ever controversial, yet essential, part of human-in-the-loop tracking experiment data analysis. Most importantly, it is used to test hypotheses about changes in population parameters, where statistical models are used to assess the likelihood that a certain hypothesis is true. Many different statistical models are available to researchers to analyze differences among group means in a sample, all relying on different assumptions and having different advantages and disadvantages. The field of statistics is complex and most researchers outside of this field do not have detailed knowledge about the ever-changing state-of-the-art in statistical modeling. Often, the choice of statistical method is based on prior research in the field without careful consideration of all the available options.

Statistical analysis on human-in-the-loop experiment data can be challenging, for a number of reasons. First, a characteristic of most human-in-the-loop studies is that the sample size is generally modest and that all too often also data for some conditions are not available for some participants, e.g., due to data corruption or participant dropout. In addition, due to the fact that it is often not possible (e.g., to avoid experiments becoming impractically long in duration) to test a complete factorial set of experiment conditions, in many cases researchers compromise for a statistically incomplete design. Finally, key dependent measuresthat are often used for comparing across different conditions (e.g., performance metrics, crossover frequency) typically do not provide well-distributed (i.e., normal) samples, due to large between-subject differences or sub-populations in the subject pool (e.g., high/low-gain pilots [1–3]). Despite these typical issues, the default choice for statistical analysis in most human-in-the-loop simulation studies is still to perform a (repeated-measures) analysis of variance (ANOVA), or a robust alternative [1, 2, 4–11].

Linear mixed-effects models or, more simply, mixed models are a type of regression model that takes into account both the data variation that is explained by the independent variables of an experiment, the *fixed effects*, and the variation that is not explained by the independent variables, the *random effects* [12]. The random effects essentially provide structure to the error term of the statistical model. Mixed models are widely used in the fields of ecology and biology [13], which often deal with complex and statistically messy datasets. Mixed models have many advantages over more traditional statistical methods like the ANOVA, such as the way missing data are handled and their applicability to more complex data structures [12]. Due to these advantages, it is surprising that mixed models are not more widely considered for experimental human-in-the-loop studies.

This paper tests the effectiveness of linear mixed-effects models by re-analyzing tracking data from previous human-in-the-loop tracking experiments. On purpose, we analyze data from two separate yet closely related roll tracking experiments [1, 9, 14] performed at TU Delft. In Experiment 1 [1], seven Cessna Citation II pilots performed the same compensatory roll attitude tracking task both in real flight using TU Delft's PH-LAB laboratory aircraft, as well as in TU Delft's SIMONA Research Simulator (SRS) under four different simulator motion cueing configurations. Experiment 2 [9, 14] was a follow-up experiment with 10 experimental conditions, where the same roll tracking task was performed with a factorial variation in roll motion filter gain (1.0, 0.75, or 0.5) and break frequency (0.0, 0.5, or 1.0 rad/s), in addition to a reference no-motion condition. Experiment 2 was performed by six of the seven Citation pilots who also participated in Experiment 1. Furthermore, all simulator conditions of Experiment 1 were also tested in Experiment 2. Individually, both experiments have a statistically messy setup, with an incomplete factorial design and often awkwardly distributed samples, which limited the quality of the statistical analysis provided in the original publications [1, 9]. In combination, the data from both experiments suffer from missing participant data (Experiment 2) and missing coverage of test conditions (Experiment 1: missing many motion settings, Experiment 2: missing in-flight condition).

This paper will add to the literature as follows: 1) the benefits of linear mixed-effects models will be discussed in the context of human-in-the-loop tracking experiments, 2) we will show how the data from different experiments can be analyzed *as a single dataset* using mixed models, to gain new insights into pilot control behavior discrepancies between real and simulated flight, and 3) the mixed-model results will be directly compared to our previous statistical results as published in the original publications [1, 9].

This paper is structured as follows. First, Section II will give an overview of linear mixed-effects models including their main assumptions. Next, Section III will provide the relevant details regarding the experiments of [1, 9], followed by the mixed-model analysis results and discussion in Section IV. The paper ends with conclusions.

II. Linear Mixed-Effects Models

A. Definition

Linear mixed-effects models (LMM) are statistical models that incorporate both fixed effects and random effects. A mixed model can be represented in matrix notation by:

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{\epsilon} \tag{1}$$

with $\boldsymbol{y}$ a known vector of observations, $\boldsymbol{\beta}$ an unknown vector of fixed effects, $\boldsymbol{u}$ an unknown vector of random effects, and $\boldsymbol{\epsilon}$ an unknown vector of random errors. X and Z are known design matrices relating the observations $\boldsymbol{y}$ to $\boldsymbol{\beta}$ and $\boldsymbol{u}$, respectively.

The two most commonly used approaches to parameter estimation in linear mixed-effects models are maximum likelihood and restricted maximum likelihood methods. When used for statistical inference, mixed-effects models of the form of Eq. (1) are progressively built-up by adding different candidate fixed effects $\boldsymbol{\beta}$ and their interactions one-by-one, starting with an intercept-only (no variation across conditions) model. Likelihood ratio tests between models with and without a candidate fixed effect are then used to determine the significance of each fixed effect in explaining the observations $\boldsymbol{y}$.

B. Assumptions

When using linear mixed-effects models for statistical analysis of effects present in a measured set of data, the following main assumptions apply [15]:

1) Linearity: the observations $\boldsymbol{y}$ are a linear combination of $\boldsymbol{\beta}$ and $\boldsymbol{u}$.
2) Absence of collinearity: the fixed effects $\boldsymbol{\beta}$ are not correlated with each other.
3) Homoskedasticity: the variance of the data should be approximately equal across the range of predicted values.
4) Normality of residuals: the model residuals $\boldsymbol{\epsilon}$ should be (approximately) normally distributed.

The normality of residuals assumption is considered to be the least important, as LMM are relatively robust against violations of this assumption. All these assumptions can be evaluated using scatter and Q-Q plots of the residuals.

C. Application to Tracking Experiment Data

This paper uses R, a free programming language and software environment for statistical computing, and the lme4 package to estimate the linear mixed-effects model parameters [12]. In the context of the experiments discussed in this paper, mixed models will be identified to describe pilot performance and control activity, and the estimated parameters of multimodal pilot models (i.e., the observations $\boldsymbol{y}$). The experimental conditions, that is, the different motion configurations and experiment environments [1, 7, 16], will be the fixed effects of the model. In addition, the interactions of the main fixed effects will be included in the models. For a within-subject (repeated measures) human-in-the-loop experiment, a between-subject "pilot" effect is the only random effect included in the models.

III. Human-in-the-Loop Experiment Datasets

This paper combines data from two experiments. Both experiments used the same tracking task and simulator setup. Experiment 1 is discussed in detail in [1] and Experiment 2 in [9, 14].

A. Tracking Task

Fig. 1 shows the compensatory roll attitude target-following disturbance-rejection task used both experiments. Pilots actively minimized the deviation of the aircraft roll attitude (ϕ) from a desired roll attitude, as defined by the forcing function f_t. In addition, pilots simultaneously counteracted a disturbance acting on the aircraft, induced by a disturbance forcing function f_d. Both the target and disturbance forcing function signals were multisines, for details

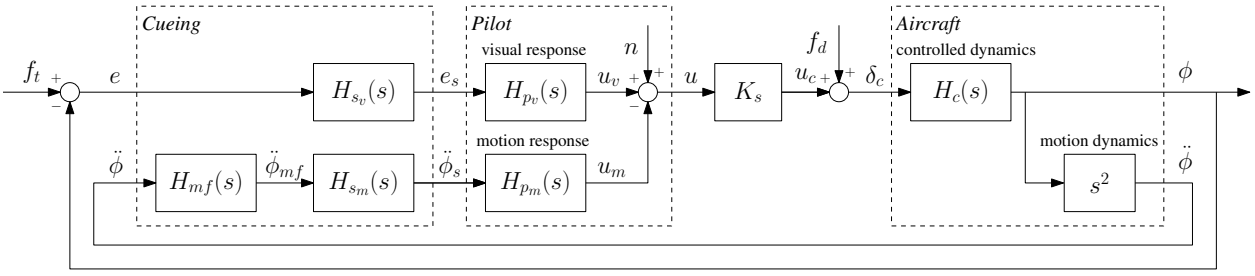


Fig. 1 Compensatory tracking task from [1, 9].

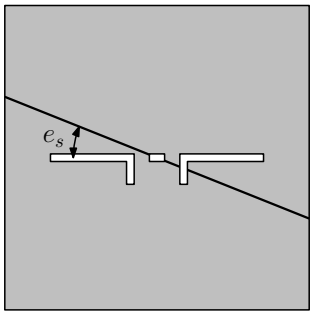


Fig. 2 Compensatory display from [1, 9].

please refer to [1]. Deviations from the target forcing function were visually presented as the tracking error e on a compensatory display, see Fig. 2.

In the tracking task, the pilots controlled the Cessna Citation II roll dynamics $H_c(s)$ using a sidestick. The summation of the disturbance forcing function f_d and the stick output u served as the control input. Furthermore, depending on the experimental condition, the pilots experienced physical motion cues as provided by the simulator motion system or the movement of the real aircraft in the in-flight condition. As shown in Fig. 1, for this control task, pilots' control dynamics can be modeled using linear visual ($H_{p_v}(s)$) and motion ($H_{p_m}(s)$) response functions and a remnant signal n accounting for nonlinear behavior and noise.

Finally, the main goal of the experiments of [1, 9, 14] was to explicitly measure how pilots' control behavior was affected by variation in simulator roll motion cueing. In both experiments, the tested variation consisted of different parameter settings for the roll motion filter, indicated with $H_{mf}(s)$ in Fig. 1. Note that the visual and motion cueing dynamics H_{s_v} and H_{s_m} (i.e., cueing delays) were also explicitly accounted for and, whenever possible, made similar between the in-flight and simulator measurements [1].

B. Experimental Conditions

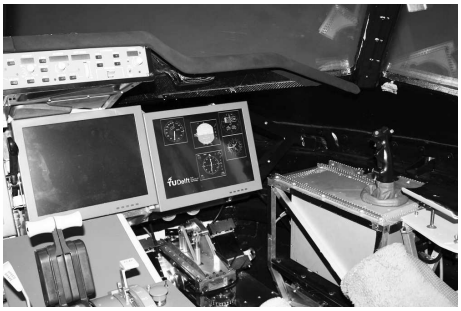
Experiment 1, as described in full detail in [1], performed a direct comparison of roll tracking behavior measured both in real flight, using TU Delft's Cessna Citation II laboratory aircraft, see Fig. 3a, and in the SIMONA Research Simulator (SRS), see Fig. 3c. To be able to perform the tracking task in real flight, the aircraft was equipped with a custom FBW control system [1, 16, 17]. Pilots were seated in the right-hand pilot seat and performed the task with a right-handed sidestick (a force stick). To ensure a fair comparison, the experiment setup in the SRS cockpit was matched to the setup in the Citation cockpit, see Fig. 3b and 3d. Experiment 2 [9, 14] made use of exactly the same experiment setup in the SRS as used for Experiment 1 (Figs. 3c and 3d) and tested a much larger set of (factorial) roll motion cueing gain and break frequency variations, see Table 1.

Table 1 lists the experimental conditions tested in Experiments 1 [1] and 2 [9, 14]. Compared to the original publications, we use a different naming convention for the different test conditions in this paper. The first character of all condition names is either a "C" for "Citation", which indicates a condition performed in the real aircraft, or an "S" for "SRS", i.e., a simulator condition. The remainder of the condition names are composed of the tested combination of roll motion filter gain K_{mf} and break frequency ω_{mf} (separated by a slash symbol), between brackets. Hence, for example, "C(1.0/0.0)" refers to the in-flight condition with 1-to-1 aircraft roll motion, while "S(0.5/1.0)" tested a roll motion filter with K_{mf} = 0.5 and ω_{mf} = 1.0 rad/s in the SRS.

(a) The Cessna Citation II laboratory aircraft.

(b) The Cessna Citation II cockpit setup.

(c) The SIMONA Research Simulator (SRS).



(d) The SRS cockpit setup.

Fig. 3 TU Delft's Cessna Citation II laboratory aircraft (PH-LAB) and SIMONA Research Simulator and their cockpits during the experiments of [1, 9, 14].

Table 1 Roll task experimental conditions for Experiments 1 and 2.

Condition	Apparatus	K_{mf}	ω_{mf}	Description	Experiment 1	Experiment 2
S(0.0/-)	SRS	0.0	–	no motion	√	√
S(0.5/1.0)	SRS		1.0 rad/s	strong washout, low gain		√
S(0.5/0.5)	SRS	0.5	0.5 rad/s	medium washout, low gain	√	√
S(0.5/0.0)	SRS		0.0 rad/s	no washout, low gain		√
S(0.75/1.0)	SRS		1.0 rad/s	strong washout, medium gain		√
S(0.75/0.5)	SRS	0.75	0.5 rad/s	medium washout, medium gain		√
S(0.75/0.0)	SRS		0.0 rad/s	no washout, medium gain		√
S(1.0/1.0)	SRS		1.0 rad/s	strong washout, high gain		√
S(1.0/0.5)	SRS	1.0	0.5 rad/s	medium washout, high gain	√	√
S(1.0/0.0)	SRS		0.0 rad/s	no washout, high gain	√	√
C(1.0,0.0)	Citation	1.0	0.0 rad/s	no washout, high gain	√	

Table 1 shows that all simulator conditions tested in Experiment 1 were also part of the, more complete and elaborate, test condition matrix of Experiment 2. The reference in-flight condition C(1.0/0.0) was, however, only evaluated in Experiment 1.

C. Participants

Experiments 1 and 2 were both performed by a group of Cessna Citation pilots. Experiment 1 was performed by a total of seven pilots, of whom six returned for Experiment 2. As in this paper the data from both experiments are combined and differences between individual pilots are explicitly accounted for in our mixed models, Table 2 lists the mapping of pilot/subject numbers used in [1, 9, 14].

D. Dependent Measures and Data Analysis

For both Experiments 1 and 2, a number of different dependent measures were used for comparing pilots' control behavior and task performance across the tested conditions. The following different categories of dependent measures are presented in [1, 9, 14]:

Table 2 Mapping of pilot/subject identification numbers in [1] and [9, 14].

Source	Pilot/Subject Number						
Experiment 1 [1]	1	2	3	4	5	6	7
Experiment 2 [9, 14]	6	4	5	2	n/a	3	1

- *Task performance metrics*: the variances of the error signal e (see Fig. 1) and the control signal u were used as metrics of task performance (tracking accuracy) and control activity, respectively. In [1, 9] also the separate contributions of f_t, f_d, and n (see Fig. 1) were compared across conditions, in addition to the total variances (σ_e^2 and σ_u^2).
- *Crossover characteristics*: the crossover frequencies and phase margins for the "target" open-loop system (i.e., $\omega_{c,t}$ and $\varphi_{m,t}$) and the "disturbance" open-loop system (i.e., $\omega_{c,d}$ and $\varphi_{m,d}$) were computed. Due to the fact that Experiments 1 and 2 used a combined target-following and disturbance-rejection task, crossover characteristics were evaluated separately for both loops.
- *Pilot model parameters*: fitted linear models for pilots' visual (H_{p_v}) and motion (H_{p_m}) responses (see Fig. 1) were used to explicitly quantify pilots' multi-channel control dynamics. In both experiments, the fitted pilot model was:

$$H_{p_v}(s) = K_v\,(T_L s + 1)\,e^{-s\tau_v}\,\frac{\omega_{nm}^2}{s^2 + 2\zeta_{nm}\omega_{nm}s + \omega_{nm}^2} \tag{2}$$

$$H_{p_m}(s) = K_m\frac{5.97\,(0.11s+1)}{(5.9s+1)\,(0.005s+1)}e^{-s\tau_m}\,\frac{\omega_{nm}^2}{s^2 + 2\zeta_{nm}\omega_{nm}s + \omega_{nm}^2} \tag{3}$$

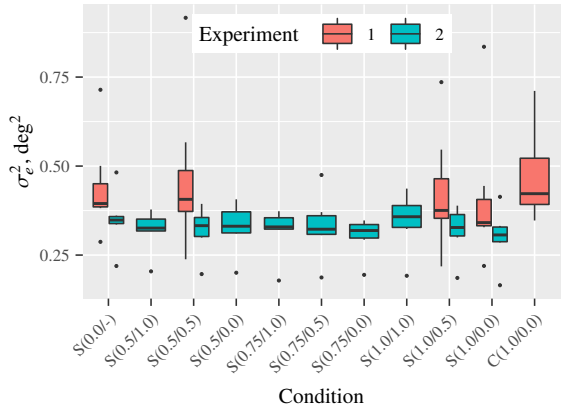
  The pilot model given by Eq. (2) and (3) had 7 free parameters: a pilot visual gain K_v, a pilot motion gain K_m, a pilot visual lead time constant T_L, visual and motion time delays τ_v and τ_m, and a neuromuscular system natural frequency (ω_{nm}) and damping ratio (ζ_{nm}). The first second-order transfer function in Eq. (3) accounts for the vestibular sensor (semi-circular canal) dynamics [1]. Finally, to objectively quantify the relative contributions of H_{p_v} and H_{p_m} to pilots' control inputs u, the fitted models were also used to simulate u_v and u_m (see Fig. 1) and calculate a relative visual/motion control signal variance fraction, $\sigma_{u_m}^2/\sigma_{u_v}^2$.

In this paper, we will analyze only the data from a subset of these original dependent measures, focusing on a number of the key metrics revealing how pilots were affected by motion cueing variations as concluded in [1, 9, 14]. Here, we will focus on using mixed-effects models to re-analyze the data from Experiments 1 and 2 for the error signal variance (σ_e^2), the control signal variance (σ_u^2), the pilot visual gain (K_v), the pilot visual lead time constant (T_L), and the visual/motion variance fraction $\sigma_{u_m}^2/\sigma_{u_v}^2$. Fig. 4 to 6 show the combined data from Experiment 1 (red) and Experiment 2 (green) for all experiment conditions.

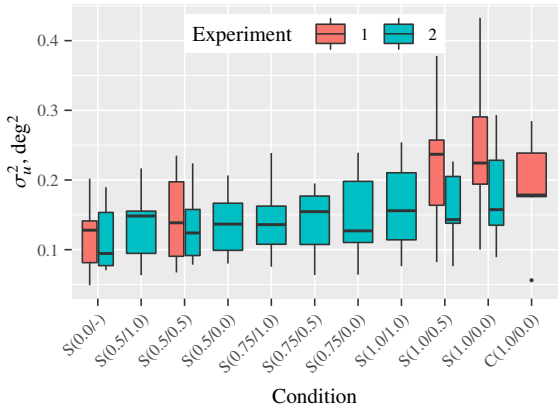
For statistical comparisons of the Experiment 1 and Experiment 2 data across the tested conditions, Analysis of Variance (ANOVA) tests were used in [1, 9, 14]. For Experiment 1, which did not have a factorial set of test conditions defined by multiple independent variables, see Table 1, a 1-way repeated measures ANOVA was used as the main statistical test [1]. In cases where data distributions were strongly non-normal (e.g., σ_e^2), a nonparametric Friedman test was applied. Experiment 2 tested a complete factorial combination of three K_{mf} and three ω_{mf} settings, in addition to a reference no-motion condition S(0.0/-). The original statistical analysis for Experiment 2 [9, 14] was performed using a 2-way factorial repeated measures ANOVA on all conditions with simulator motion, with K_{mf} and ω_{mf} as statistical factors, hence excluding the no-motion case. Overall, the statistical comparisons of the data reported in [1, 9, 14] were suboptimal, due to problems with data distributions (e.g., σ_e^2 in Experiment 1) and the fact that the tested conditions were not a balanced factorial combination of independent variables (Experiments 1 and 2).

IV. Results

This section presents the results of the improved statistical analysis of the measured data from Experiments 1 and 2 combined, as shown in Fig. 4 to 6, using mixed-effects models. In addition to exploring the possibilities for performing mixed-model analysis on the combined dataset, we will first use Experiment 2's data for a direct comparison of basic statistical outcomes between mixed-effects models and ANOVAs. To conclude, we provide a reflection on the merits

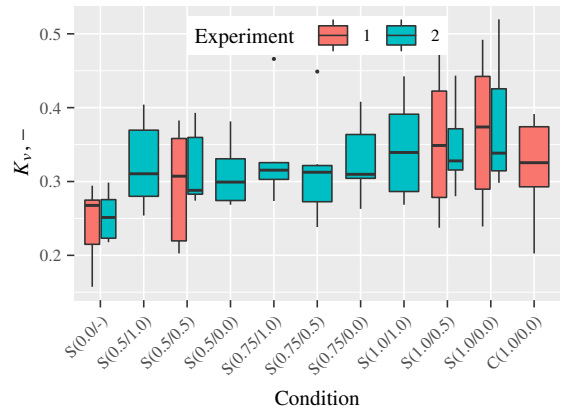


(a) Error variance.

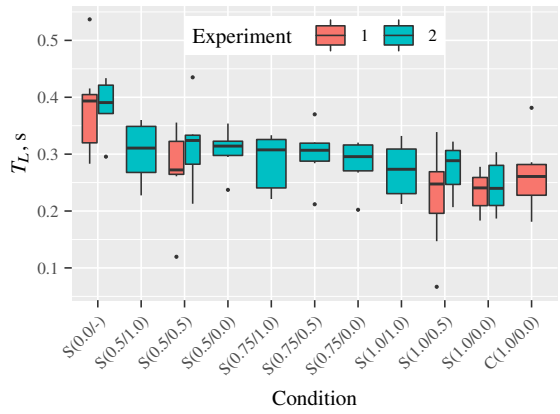


(b) Control input variance.

Fig. 4 Pilot performance and control activity.



(a) Visual gain.



(b) Lead time constant.

Fig. 5 Pilot equalization parameters.

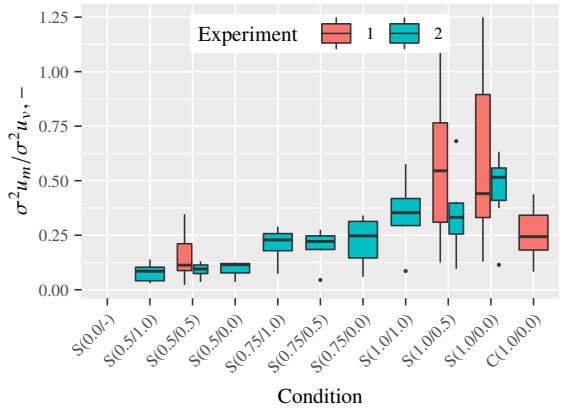


Fig. 6 Visual/motion variance fraction.

of mixed-effect models for human-in-the-loop experiment data analysis and the main statistical conclusions regarding the statistical comparisons (e.g., in-flight vs. simulator) made in [1, 9, 14].

Some outliers were present in the data, as assessed by boxplots (see Fig. 4 to 6), but they were kept in for the analysis. Assumptions of linearity, homoskedasticity, and normality of residuals were checked visually using scatter plots and Q-Q plots of the residuals for each model. For all dependent measures, except the lead time constant T_L, the model assumptions of homoskedasticity and normality of residuals were not met due to the fact that these measures were positively skewed. These dependent measures were transformed to make the data more normally distributed using Box-Cox transformations [18] with λ values of -0.18, 0.18, 0.1, and 0.14 for σ_e^2, σ_u^2, $\sigma_{u_m}^2/\sigma_{u_v}^2$, and K_v, respectively. All models met the assumption of homoskedasticity and normality of residuals after the transformations were applied. No other violations of the assumptions were detected.

A. Mixed-Effects Models vs. ANOVA (Experiment 2)

In [9, 14], the original statistical analysis for Experiment 2 was performed using a two-way repeated-measures ANOVA on the data from all motion conditions only (so excluding the no-motion condition S(0.0/-)). The two statistical factors that varied across these conditions were the motion filter gain K_{mf} and break frequency ω_{mf}, both with three levels. Here, we performed a direct comparison of theses originally reported statistical test results with the outcomes of a linear mixed-effects model (LMM) analysis on the same data. As explained in Section II, the LMM used for this analysis considers K_{mf} and ω_{mf} as the fixed effects, while the between-subject differences are accounted for as a random effect.

Table 3 shows a direct comparison of the LMM and ANOVA analysis for the five dependent measures compared in this paper: σ_e^2, σ_u^2, K_v, T_L, and $\sigma_{u_m}^2/\sigma_{u_v}^2$. For both tests, the significance of the direct effects of K_{mf} and ω_{mf} was tested, as well as their interaction $K_{mf} \times \omega_{mf}$. For the LMM analysis, we report the χ^2 value for addition of each factor with the degrees-of-freedom in brackets (df), as well as the statistical significance. Consistent with [9, 14], we report the effect and error degrees-of-freedom df, the F test statistic, and the significance of each factor for the ANOVA.

Table 3 Comparison of LMM and ANOVA results for data from [9, 14].

Dependent Measures	Linear Mixed-Effects Models						Repeated-Measures ANOVA								
	Factors						Factors								
	K_{mf}		ω_{mf}		$K_{mf} \times \omega_{mf}$		K_{mf}			ω_{mf}			$K_{mf} \times \omega_{mf}$		
	$\chi^2(2)$	Sig.	$\chi^2(2)$	Sig.	$\chi^2(4)$	Sig.	df	F	Sig.	df	F	Sig.	df	F	Sig.
σ_e^2	0.52	–	4.89	*	14.12	**	2,10	0.41	–	2,10	4.40	**	4,20	2.67	*
σ_u^2	18.43	**	2.63	–	1.58	–	2,10	10.03	**	2,10	1.00	–	4,20	0.37	–
K_v	18.93	**	1.90	–	7.26	–	2,10	10.87	**	2,10	3.10	*	4,20	1.25	–
T_L	20.14	**	4.83	*	2.35	–	2,10	6.74	**	2,10	4.53	**	4,20	0.59	–
$\sigma_{u_m}^2/\sigma_{u_e}^2$	58.82	**	3.87	–	3.75	–	2,10	20.54	**	2,10	6.03	**	4,20	1.70	–

** = significant ($p < 0.05$)
* = weakly significant ($0.05 \leq p < 0.1$)
– = not significant ($p \geq 0.05$)

As expected, Table 3 shows that the statistical analysis outcomes of both approaches are very similar. Especially equivalent statistically significant effects of K_{mf} (σ_u^2, K_v, T_L, and $\sigma_{u_m}^2/\sigma_{u_v}^2$) and $K_{mf} \times \omega_{mf}$ (σ_e^2) are found for both the LMM and the ANOVA across all considered dependent measures. Overall, somewhat weaker statistical effects of ω_{mf} were found with the LMM, i.e., less significant effects were found than with the ANOVA (K_v and $\sigma_{u_m}^2/\sigma_{u_v}^2$), as well as higher p-values for the significant effects that were found (σ_e^2 and T_L). In [9, 14] it is concluded that the effects of varying K_{mf} on the pilots' manual control behavior were stronger than those of ω_{mf}. This can also be confirmed from visual inspection of the data in Fig. 4 to 6. Thus, the subtle differences between the LMM and ANOVA outcomes listed in Table 3 are explained by the fact that the not very strong effects of ω_{mf} in this (small) 6-pilot dataset are likely to result in somewhat inconsistent statistical outcomes.

B. Mixed-Effects Model Analysis on Combined Dataset

Table 4 provides the results of the linear mixed effects model analysis of the combined dataset from the two experiments and for the five dependent measures considered in this paper. The models included motion filter gain

(K_{mf}), motion filter break frequency (ω_{mf}), apparatus (APP), and experiment (EXP) as fixed effects. In addition, the interaction of K_{mf} and ω_{mf}, as well as the interactions of EXP with both K_{mf} and ω_{mf}, were included as additional fixed effects. Pilot was used as the random effect. Even though linear mixed effects models can consist of a mix of categorical and continuous variables, both K_{mf} and ω_{mf} were treated as categorical as this is more in line with the ANOVA analysis and ω_{mf} does not apply to the no-motion condition S(0.0/-). The models included random intercepts only, i.e., no random slopes were introduced.

The mixed-effects models were progressively built up by adding the different fixed effects and interactions one-by-one, starting with the intercept-only model. Likelihood ratio tests between the models with and without a fixed effect or interaction determined the significance of that effect. The fixed factors were added to the model in the order from left to right in Table 4. To illustrate the fit of different models, Fig. 7 provides predictions of some of these different models for the pilot control activity dependent measure σ_u^2. For clarity, this figure only includes the raw data for the simulator conditions (see Table 1) and the corresponding mixed-effects model predictions. Fig. 7a shows the intercept-only model. This is the most basic model of the data corresponding to the mean. Fig. 7b shows the model with the apparatus (APP) and motion gain (K_{mf}) factors included as fixed effects. Both these factors in this case are significant, see Table 4. Note that, visually this model indeed also provides a better fit to the data. Thereby, the model thus reveals a significant increase in pilot control activity when the motion gain increases. Finally, Fig. 7c provides the final model including all fixed effects listed in Table 4. Some variation in predictions can be observed from the motion filter break frequency (ω_{mf}), however, as Table 4 shows, this added variation is not a significant effect. The effects of experiment (EXP) and the interaction between experiment and motion gain (K_{mf}) are significant and reflect the overall higher control activity and steeper increase of pilot control activity in Experiment 1 compared to Experiment 2.

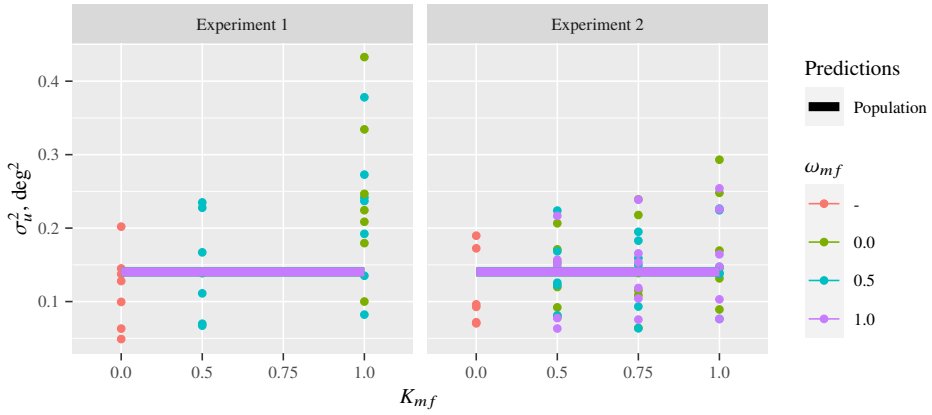
Table 4 Linear mixed effects model comparison statistics for combined data set.

	Factors													
Dependent	APP		K_{mf}		ω_{mf}		$K_{mf} \times \omega_{mf}$		EXP		EXP $\times$ K_{mf}		EXP $\times$ ω_{mf}	
Measures	$\chi^2(1)$	Sig.	$\chi^2(3)$	Sig.	$\chi^2(2)$	Sig.	$\chi^2(4)$	Sig.	$\chi^2(1)$	Sig.	$\chi^2(2)$	Sig.	$\chi^2(1)$	Sig.
σ_e^2	16.49	**	12.22	**	5.20	*	1.67	–	37.30	**	1.48	–	0.03	–
σ_u^2	4.46	**	40.75	**	3.77	–	4.23	–	17.80	**	6.26	**	0.01	–
K_v	0.48	–	76.90	**	2.06	–	4.43	–	0.25	–	1.90	–	0.33	–
T_L	1.40	–	61.22	**	1.04	–	2.41	–	5.87	**	1.48	–	1.50	–
$\sigma_{u_m}^2/\sigma_{u_e}^2$	0.10	–	[a]87.79	**	3.36	–	3.19	–	5.66	**	0.00	–	0.81	–

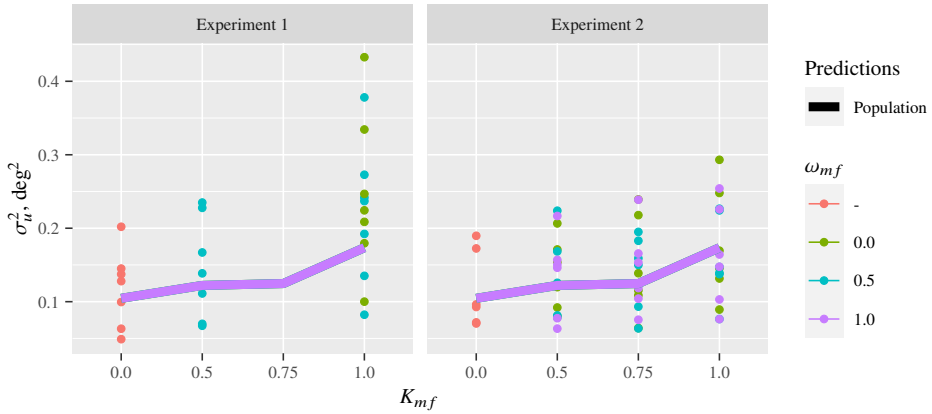
** = significant ($p < 0.05$)
* = weakly significant ($0.05 \leq p < 0.1$)
– = not significant ($p \geq 0.05$)
[a] = df of 2 due to missing data for C(0.0/0.0)

Summarizing the results from Table 4 and correlating them with Figs. 4 to 6, the following observations can be made. Pilot tracking performance was significantly worse and control activity significantly higher in the aircraft compared to the simulator (APP factor). In addition, pilots used significantly more motion in the aircraft compared to the simulator (higher $\sigma_{u_m}^2/\sigma_{u_v}^2$). The pilot visual gain and lead time constant were not affected by the apparatus. The motion gain introduced significant differences in all dependent measures considered in this paper. When the motion gain increased, tracking performance decreased, control activity increased, pilots' use of motion increased, and the pilot visual gain increased and lead time constant decreased significantly. The motion filter break frequency introduced a weakly significant effect on σ_e^2 only, as tracking performance degraded slightly with increasing ω_{mf}. The interaction between motion filter gain and break frequency was not significant for any of the dependent measures.

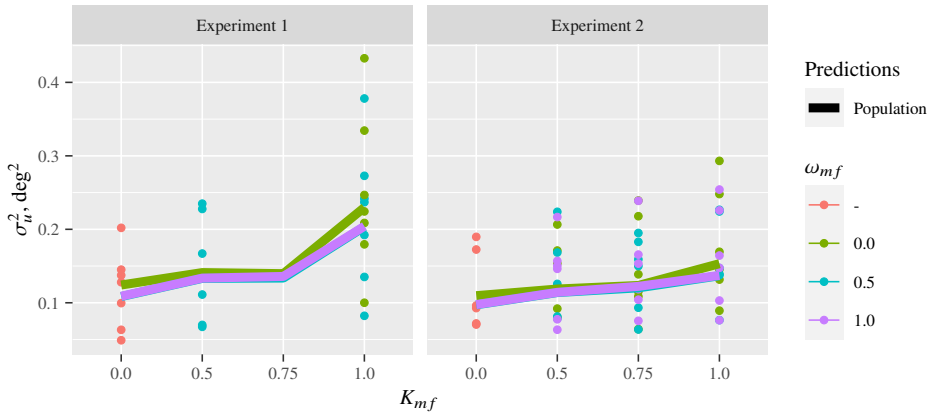
Pilot performance and control activity were also found to be significantly different between Experiments 1 and 2 (EXP factor in Table 4). Both σ_e^2 and σ_u^2 were significantly higher in the first experiment compared to the second experiment. The visual/motion variance fraction was significantly higher in the first experiment. In addition, even though a less pronounced effect, as can be verified from Fig. 5b, the pilot lead time constant was significantly lower in Experiment 1 compared to Experiment 2. Finally, Table 4 shows the results for the interactions of our EXP factor with the variations in the metrics due to K_{mf} and ω_{mf}. A significant result for these interactions indicates that the magnitudes of the change across conditions due to the motion filter gain and break frequency, respectively, were different between both experiments. Here, we find only one significant interaction with our EXP factor, and hence consistent effects of K_{mf} and ω_{mf} across the two experiments. The EXP $\times$ K_{mf} factor introduced a significant effect for σ_u^2 only, which is indeed consistent with the stronger increase in control activity with increasing K_{mf} found in Experiment 1, see Figs. 4b and 7c.



(a) Intercept only model.



(b) Model with apparatus and motion gain as factors.



(c) Final model.

Fig. 7 Pilot control activity model predictions.

C. Reflection on Experiment 1 and 2 Statistical Analysis

Overall, the mixed-effects model analysis performed on the combined data from Experiments 1 and 2 results in very similar statistical outcomes, as summarized in Table 4, as presented in the original publications [1, 9, 14]. Most of the significant effects of K_{mf}, ω_{mf}, and their interaction reported in [9, 14] for Experiment 2 (which tested a total of nine combinations of both parameters) on the dependent measures considered here are also significant from the

mixed-effects model analysis performed on both experiments' data combined. This, for example, holds for the highly significant effects of K_{mf} on all dependent measures except σ_e^2, as can be verified from comparing Tables 3 and 4. For the combined analysis, however, Table 4 also shows a significant effect K_{mf} on tracking performance (σ_e^2), which was not reported in [9]. As is shown in Fig. 4a, the effect of K_{mf} variations on σ_e^2 across conditions S(0.0/-) and S(1.0/0.0), as well as S(0.5/0.5) and S(1.0/0.5), was stronger in Experiment 1 compared to Experiment 2, which thus explains this additional significant effect. Similarly, the significant effect of ω_{mf} on T_L reported for Experiment 2 (see Table 3) is no longer significant in the combined analysis due to the lack of variation between different ω_{mf} settings in Experiment 1. Thus, overall, the combined analysis of the effects of K_{mf} and ω_{mf} in the data from this set of two experiments has only resulted in minor differences in significance for effects that were, in the original publications, concluded to be comparatively weak. On the other hand, it has not changed, but only confirmed, the main findings and conclusions regarding the effects of K_{mf} and ω_{mf} variations reported in [1, 9, 14].

One distinct benefit of using mixed-effects models for statistical analysis on a combined dataset, compared to using a traditional approach (i.e., ANOVA) for this, is that having only partial overlap in tested conditions and subjects between experiments (and hence many "missing values", as is the case here for Experiment 1) is not a problem. This ensures that the valuable statistical insights that can be gained from such combined analyses can be obtained for many cases where this would be otherwise impossible, such as the datasets used in this paper. Considering the results obtained for our EXP factor in Table 4, the mixed-effects model results indicate that σ_e^2, σ_u^2, and T_L vary significantly between both experiments. Indeed, Figures 4a, 4b, and 5b show consistently higher σ_e^2 (worse performance), increased σ_u^2 (more control effort), and lower T_L (less visual lead), respectively, in Experiment 1. While Experiment 2 was a standalone simulator experiment [9], Experiment 1 was performed on the same day (and after) the pilots had already performed a similar pitch tracking experiment (as reported in [7]). Hence, the effects of EXP found in Table 4 may be explained by different levels of engagement/fatigue in the task across both experiments. Furthermore, the fact that the EXP factor in our analysis only showed a significant interaction with the applied motion cueing variations in a single instance (σ_u^2 for K_{mf}) helps statistically underline the consistency of the trends observed in both experiments.

Similarly, this benefit of the mixed-models analysis is also found to facilitate improved analysis of the combined in-flight and simulator data as previously performed for Experiment 1 only [1]. In our current analysis with mixed-models, the in-flight condition C(1.0/0.0) was statistically modeled as an additional instance of a "no washout, high gain" condition, but with a potentially different intercept (i.e., offset value) accounted for with our APP variable (see Table 1). Where in [1] post-hoc tests were performed to (mostly without success) verify if any significant differences between the simulator and in-flight data existed, in our current analysis APP introduces highly significant effects for σ_e^2 and σ_u^2 that were indeed reported as key (yet statistically not always significant) differences in [1]. It should be noted that this benefit of course is not restricted to a combined experiment data analysis as presented here in Table 4, but also for mixed-effects model analysis on the data from Experiment 1 only (not performed here). Thus, especially the capacity for straightforward and flexible evaluation of the effects of different facilities in a statistical factor (our APP) can be highly valuable for especially in-flight vs. simulator comparisons.

V. Conclusions

This paper used mixed-effects models to perform a renewed and extended statistical analysis on the data from two related human-in-the-loop tracking experiments. Both experiments compared pilot tracking behavior in a roll attitude tracking task with different roll motion cueing (gain and high-pass filter break frequency) settings in a simulator, while one experiment also included an in-flight measurement for the same task. Our mixed-effects model analysis confirmed the main statistical outcomes of both individual experiments as reported in our previous publications [1, 9, 14]. Furthermore, mixed-effects models were found to facilitate statistically meaningful comparison of trends observed in multiple separate human-in-the-loop experiments (e.g., replication experiments), even in cases of only partial overlap in tested conditions. Finally, the fact that mixed-effects models can inherently cope well with missing cases and additional environment variables (i.e., in-flight vs. simulator) was found to be especially helpful for the statistical analysis of human-in-the-loop experiments as considered in this paper. Such experiments often test additional "reference" conditions (e.g., no motion, or in-flight) in addition to a (factorial) variation in independent variable(s) that often prevent the use of more traditional statistical analysis methods on the complete set of experiment data.

References

[1] Pool, D. M., Zaal, P. M. T., Damveld, H. J., van Paassen, M. M., and Mulder, M., "Evaluating Simulator Motion Fidelity using In-Flight and Simulator Measurements of Roll Tracking Behavior," *Proceedings of the AIAA Modeling and Simulation Technologies Conference 2012, Minneapolis (MN)*, 2012.

[2] Pieters, M. A., Zaal, P. M. T., Pool, D. M., Stroosma, O., and Mulder, M., "A Simulator Comparison Study into the Effects of Motion Filter Order on Pilot Control Behavior," *Proceedings of the AIAA Modeling and Simulation Technologies conference, San Diego (CA)*, 2019. https://doi.org/10.2514/6.2019-0712.

[3] Mitchell, D. G., and Klyde, D. H., "Defining Pilot Gain," *Journal of Guidance, Control, and Dynamics*, Vol. 43, No. 1, 2020, pp. 85–95. https://doi.org/10.2514/1.G004426.

[4] Pritchett, A. R., and Yankosky, L. J., "Pilot-Performed In-Trail Spacing and Merging: An Experimental Study," *Journal of Guidance, Control, and Dynamics*, Vol. 26, No. 1, 2003, pp. 143–150. https://doi.org/10.2514/2.5025.

[5] Zaychik, K. B., and Cardullo, F. M., "Simulator Sickness, Workload and Performance as a Function of Visual Delay in a Simulator," *Proceedings of the IAA Modeling and AIAA Simulation Technologies Conference and Exhibit, San Francisco (CA)*, 2005. https://doi.org/10.2514/6.2005-6302.

[6] Damveld, H. J., Beerens, G. C., van Paassen, M. M., and Mulder, M., "Design of Forcing Functions for the Identification of Human Control Behavior," *Journal of Guidance, Control, and Dynamics*, Vol. 33, No. 4, 2010, pp. 1064–1081. https://doi.org/10.2514/1.47730.

[7] Zaal, P. M. T., Pool, D. M., van Paassen, M. M., and Mulder, M., "Comparing Multimodal Pilot Pitch Control Behavior Between Simulated and Real Flight," *Journal of Guidance, Control, and Dynamics*, Vol. 35, No. 5, 2012, pp. 1456–1471. https://doi.org/10.2514/1.56268.

[8] Correia Grácio, B. J., Valente Pais, A. R., van Paassen, M. M., Mulder, M., Kelly, L. C., and Houck, J. A., "Optimal and Coherence Zone Comparison Within and Between Flight Simulators," *Journal of Aircraft*, Vol. 50, No. 2, 2013, pp. 493–507. https://doi.org/10.2514/1.C031870.

[9] Pool, D. M., van Paassen, M. M., and Mulder, M., "Effects of Motion Filter Gain and Break Frequency Variations on Pilot Roll Tracking Behavior," *Proceedings of the AIAA Modeling and Simulation Technologies Conference, Boston (MA)*, 2013.

[10] Grant, P. R., Moszczynski, G., and Schroeder, J. A., "Post-stall Flight Model Fidelity Effects on Full Stall Recovery Training," *Proceedings of the AIAA Modeling and Simulation Technologies Conference, Atlanta (GA)*, 2018. https://doi.org/10.2514/6.2018-2937.

[11] D'Intino, G., Olivari, M., Bülthoff, H. H., and Pollini, L., "Haptic Assistance for Helicopter Control Based on Pilot Intent Estimation," *Journal of Aerospace Information Systems*, Vol. 17, No. 4, 2020, pp. 193–203. https://doi.org/10.2514/1.I010773.

[12] Bates, D., Mächler, M., Bolker, B., and Walker, S., "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, Vol. 67, No. 1, 2015. https://doi.org/10.18637/jss.v067.i01.

[13] Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C., Robinson, B. S., Hodgson, D. J., and Inger, R., "A brief introduction to mixed effects modelling and multi-model inference in ecology," *PeerJ*, Vol. 6, No. e4794, 2018. https://doi.org/10.7717/peerj.4794.

[14] Pool, D. M., "Objective Evaluation of Flight Simulator Motion Cueing Fidelity Through a Cybernetic Approach," Ph.D. thesis, Delft University of Technology, Faculty of Aerospace Engineering, Sep. 2012. URL http://repository.tudelft.nl/assets/uuid:e49e4ead-22c4-4892-bbf5-5c3af46fc9f5/phdthesis_dmpool.pdf.

[15] Winter, B., "Linear models and linear mixed effects models in R with linguistic applications," , 2013. URL http://arxiv.org/pdf/1308.5499.pdf.

[16] Mulder, M., Zaal, P. M. T., Pool, D. M., Damveld, H. J., and van Paassen, M. M., "A Cybernetic Approach to Assess Simulator Fidelity: Looking back and looking forward," *Proceedings of the AIAA Modeling and Simulation Technologies Conference, Boston (MA)*, 2013.

[17] Mulder, M., Lubbers, B., Zaal, P. M. T., van Paassen, M. M., and Mulder, J. A., "Aerodynamic Hinge Moment Coefficient Estimation Using Automatic Fly-by-Wire Control Inputs," *Proceedings of the AIAA Modeling and Simulation Technologies Conference and Exhibit, Chicago (IL)*, 2009.

[18] Box, G. E. P., and Cox, D. R., "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Vol. 26, No. 2, 1964, pp. 211–252.