A Secondary Data Analysis using Bayesian Statistics to Explore the Influence of Gender and Initial Performance on Skill Acquisition using a Laparoscopy Simulator

Lielle Posen

Behavioural, Management and Social Sciences, University of Twente

1st Supervisor: Dr. Marleen Groenier

2nd Supervisor: Dr. Simone Borsci

November 25, 2020

Abstract

Background: The aim of simulators in the medical context is to move the critical part of the learning curve, where mistakes and lapses occur, from the patient to the simulator. For this to occur, selecting an optimal training strategy is necessary. For example, a proficiency-based program reduced surgery residents mistakes in their first 10 laparoscopic surgeries (Ahlberg et al., 2007). Unfortunately, current training strategies are not adapted to individual differences, which could improve effectiveness/efficiency by providing an environment for deliberate practice, where improvement occurs through conscious effort (Ericsson, 2004). Exploring individual differences would enable the development of individualized training programs and assessment procedures.
Objective: The main objective of the study is to explore how individual differences in gender and initial performance influence skill acquisition on LapSim.
The secondary objective was to use a Bayesian approach which compared to a Frequentist approach, should generate more accurate inferences as it produces better model fit for complex data.
Methods: Data was acquired by Groenier, Groenier, Miedema, & Broeders (2015) and Groenier, Schraagen, Miedema, & Broeders (2014) who used Frequentist approaches, while the current study used Bayesian. In the longitudinal study, 67 participants completed weekly 30-minute training sessions. For analysis *duration* and *damage count* assessed performance of the first 5 sessions as two tasks – *grasping and instrumental navigation* – were conducted at medium level difficulty.
Main Findings: 1) No gender differences were found for speed; however, gender differences were found for accuracy. 2) Initial performance differences were reduced with practice, for both speed and accuracy. 3) For model criticism, using gender as a level had no predictive ability, while initial performance levelling did. As gender showed no predictive ability, it would not be useful for forecasting as it does not provide additional knowledge on how participants perform. 4) For model fit, duration data showed poor fit for all distributions - ExGaussian, Gaussian, and Gamma; this poor fit may create more uncertainty and less precise estimations. Damage count data showed the best fit with a Poisson distribution.
Conclusion: No male advantage was found, which is contrary to past research where males hold an advantage for visuospatial tasks. Although females had an advantage for accuracy, it subsided with practice. As differences are not pronounced, we recommend that individualized training programs should not be implemented for gender groups; which goes against Donnon, DesCôteaux, and Violato (2005) who suggested one-on-one training was beneficial for female laparoscopic trainees.
Initial performance produces *transient* performance outcomes, as differences in initial accuracy and speed become less influential as practice occurred. From these findings, we recommend that for assessment of laparoscopic skill, one-time initial testing and screening is inappropriate and should be avoided when selecting potential trainees.
Keywords**:** *Minimally Invasive Surgery, Laparoscopy, Simulators, Individual Differences, Gender, Initial Performance, Learning Curves, Skill Acquisition, Multilevel Modelling, Bayesian Analysis*

Table of Contents

***Disclaimer**: Due to extraordinary circumstances related to the COVID-19 pandemic, it was not feasible to collect real-world data as a result of social restrictions put in place. Therefore, the current study used data collected from two previously published studies by Groenier, Groenier, Miedema, & Broeders (2015) and Groenier, Schraagen, Miedema, & Broeders (2014). While the previous research papers used traditional methods of statistical analysis, the current study reanalysed the data using Bayesian statistical methods.*

Introduction

People are fundamentally different and exploration into individual differences and how they influence the way we obtain and learn skills is one of the fundamental aspects of educational research (Donnon et al., 2005). Why do specific teaching strategies work for some people and not for others? If we were all fundamentally the same, people would learn at the same rate, understand the same instructions, and pursue the same learning strategies. However, this does not appear to be the case, and there is a need to focus research into finding specialised training programs that are made for the individual and their needs (Kolkman, Wolterbeek, & Jansen, 2005; Stefanidis, Acker, Swiderski, Heniford, & Greene, 2008). Before training programs can be implemented it is pertinent to understand the effect individual differences have on skill acquisition.

It is crucial to determine if a gender difference is apparent in surgical performance for minimally invasive surgery (MIS). There is a need for research into gender-based differences as stated in published reports by the Institute of Medicine (IOM) in 2001 and 2012 (Becker et al., 2007). The current study adds to this area by exploring gender in terms of surgical training and skill acquisition of highly visuospatial tasks. Gender differences that typically favour male participants have been previously established regarding visuospatial ability and speed tasks (Donnon et al., 2005; Thorson, Kelly, Forse, & Turaga, 2011).

Initial performance as an individual difference is often used for assessment purposes to differentiate good and bad performers. Nevertheless, previous findings have suggested that

participants with the same initial proficiency in a task may differ in later performance (Bahrick, Bahrick, Bahrick, & Bahrick, 1993). The inference from these findings is that initial performance does not necessarily determine later performance. This is especially crucial, as it indicates that one-time testing of initial performance may not be a fair representation of a candidate's future performance and abilities.

When examining individual differences it is important to explore if a specific individual difference leads to performance differences which are transient – where differences are seen only during early stages of learning; or enduring performance differences – whereby an individual difference still influences performance even after practice (Keehner, Lippa, Montello, Tendick, & Hegarty, 2006). Although, the goal is to find individual differences that cause enduring differences as they can be used for pre-screening potential candidates (Keehner et al., 2006). Past research has leaned towards the finding that most individual differences lead to transient performance outcomes (Ackerman, 1992; Keehner et al., 2006). Nevertheless, confirmation that a specific individual difference generates transient performance differences within a field is still useful information, as it confirms that such a difference should not be used for initial screenings.

The current study modelled learning curves to explore skill acquisition, as participants used a simulator known as LapSim. Simulators are a powerful research tool as they can assess skill acquisition. This allows exploration of learning curves which can determine the influence individual differences, such as *gender* and *initial performance*, have on skill acquisition at different levels of expertise (from a novice to expert); as well as determine if gender and initial performance produce performance differences which are *transient* or *enduring*.

1. Minimally Invasive Surgery (MIS)

Minimally invasive surgery (MIS) is a relatively recent breakthrough in the medical field, also known as "keyhole surgery," as small incisions are made and specialized tools with in-built cameras are used to investigate and rectify a particular internal medical problem in a patient. Overall, it has advantages in terms of the prospective patient outcome as there is reduced blood loss and faster

recovery times (Bissolati, Orsenigo, & Staudacher, 2016; Galaal et al., 2012). Laparoscopy is a form of MIS which is performed through the abdomen. However, the learning curve needed to acquire MIS skills is more prolonged compared to open surgery (Bennett, Stryker, Rosario Ferreira, Adams, & Bert, 1997). Difficulties are more pronounced: firstly, surgeons only see what is happening indirectly through 2D camera recordings. This is different to open surgery where surgeons work in a 3D environment, and the working area can directly be viewed (Perez-Cruet, Fessler, & Perin, 2002). Secondly, there is less tactile feedback (Perez-Cruet et al., 2002). Thirdly, the surgery is a complex motor task which is bimanual whereby both hands are needed. Bimanual coordination is necessary, which is the "synergistic movement of two different instruments, as well as smoothness of movements" (Rieder et al., 2011). Lastly, surgeons must account for the "fulcrum effect", whereby movements in one direction will be outputted as a movement in the opposite direction (Gallagher, McClure, McGuigan, Ritchie, & Sheehy, 1998).

1.2 Need for Simulators

While MIS has benefits, training strategies and methods are crucial for implementing it safely on patients. In terms of surgical training, simulators attempt to create an environment where skills and techniques that are learnt indirectly can be applied later in actual practice within an operating room environment.

The need for simulators in the medical field is ever-present as a way "to 'train out' the learning curve", whereby technical skills are practiced and learnt on the simulator rather than on patients (Gallagher et al., 2005). This should help with workload and attention demands during actual surgery, as these technical skills (e.g. psychomotor and spatial skills) are trained and become automatic. This leaves more attentional resources available to handle complications that may arise during actual surgery (Gallagher et al., 2005). Research by Seymour et al. (2002) shows evidence for the benefits of simulators and this ability to move the critical part of the learning curve from the patient to the simulator. In their study, residency students that used virtual reality simulator training were both faster and produced less errors during actual laparoscopic surgery, compared to those

that did not train their technical skills using a simulator. Further evidence by Ahlberg et al. (2007) found that, although simulator training is important, the type of training strategy used on the simulator is also of equal importance. They found that proficiency based simulator programs as a training method reduced laparoscopic errors for the first 10 surgeries performed by medical residency students. For this reason, it is essential to determine individual differences influencing skill acquisition. As firstly, this information can help create individualised training methods and strategies that can assist candidates in achieving competency and move errors and risks from the patient to the simulator. Secondly, these individual differences can allow for assessment procedures to determine which individuals may be better suited to perform MIS.

To explore these possibilities, simulators can be used as a useful research tool as it has the ability to measure objective technical ability, which cannot be tested directly during actual surgery (Ahlberg et al., 2007). Therefore, simulators can measure the continuous process of how skills are learnt, as well as explore in what manner individual differences may influence the learning process.

2. Performance Evaluation & Monitoring

A device can evaluate performance in order to determine if an individual has reached a sufficient and acceptable competency level (Moorthy, Munz, Sarker, & Darzi, 2003). This is known as objective assessment, and an example of this is if a person takes a test and scores 60% this may be considered as a pass while a lower score is a fail. In this case, an assessment tool, such as the test, can be defined as an instrument that allows a person to rank users and distinguish between good and bad performers (Van Dongen, Tournoij, Van Der Zee, Schijven, & Broeders, 2007).

Alternately, the monitoring of skill acquisition is focused on how an individual progresses and improves their performance and skills as they acquire additional experience (Rosser, Rosser, & Savalgi, 1997). Therefore, monitoring skill acquisition assesses the continuous process whereby an individual will start as a novice with no experience and progresses into becoming an expert. In this case, an assessment tool evaluates the process of learning skills in which individual differences may influence skill acquisition. This can help to explore whether individual differences are transient and

only crucial in the beginning or enduring and influence later performance even after practice. Additionally, monitoring how a skill is acquired can provide useful information when creating specialized training programs that focus on improving skill acquisition.

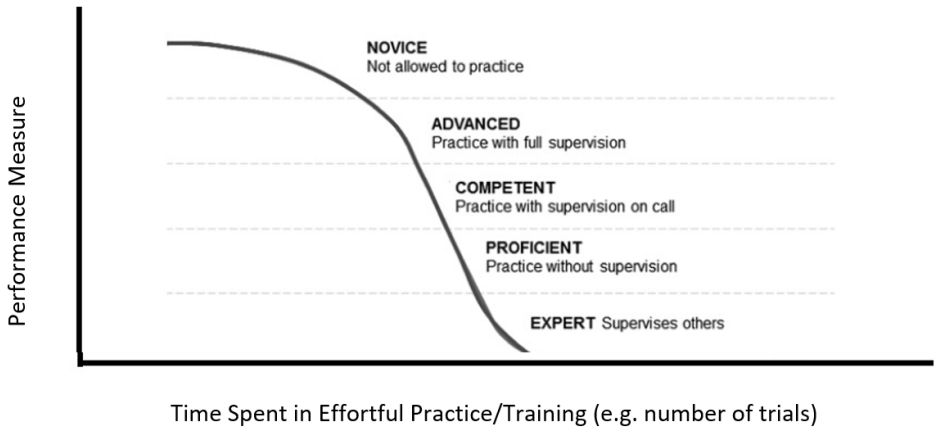
3. Skill Acquisition Models

The underlying rationale behind skill acquisition and the progression from novice to expert can be understood using the *Model of Skill Acquisition* created by Dreyfus and Dreyfus (1980). As practice takes place, a person becomes more competent in performing the skills acquired within a specific field. Throughout learning, milestones are accomplished until an individual reaches expert level proficiency. At this point extra practice does not substantially improve performance.

The Dreyfus and Dreyfus (1980) model can be used to justify the characterised shape of a typical learning curve, see Figure 1. Overall, learning curves are a useful tool to both measure and show how skills are acquired and how performance changes with time.

Figure 1

Learning Curve for Skill Acquisition



Note. The figure is an adaptation of that found in the study of Pusic et al. (2015). It shows the progression of novice to expert and the milestones, according to Dreyfus and Dreyfus (1980).

The model of skill acquisition by Ericsson (2004) proposed that to acquire a skill – and especially improve at an expert level – it is necessary for deliberate practice to take place, where a participant puts in "active effort to improve". The training method chosen is an integral aspect in producing an

environment which deliberate practice can take place, as mere exposure to a task is not sufficient (Ericsson, 2004). He notes that in the future, training devices – such as simulators – which incorporate individualized training programs, will be crucial in allowing the opportunity for learners to obtain deliberate practice and enhance skills. However, before this can be achieved, it is important to research the mechanisms (for example specific individual differences) that may be influencing performance (Ericsson, 2004).

3.1 Early vs. Late Stages of Skill Acquisition

According to Keehner et al. (2006), there are many models of skill acquisition where a switch in attention takes place depending on the stage of learning (Ackerman, 1992; J. R. Anderson, 1982; Fitts, 1964; James, 1891; Shiffrin & Schneider, 1977). During the early learning phases, where minimal practice has occurred, it is necessary for the participant to increase their cognitive attention in order to complete a task successfully. After practice takes place and repetition of the task has occurred a participant enters later stages of learning. A shift in attentional demand occurs during these later stages of learning, in which less attention is required as the task becomes more automatic and procedural. Ackerman (1992) argues that this switch in attentional demands will result in individual differences having less of an influence during later learning stages. Therefore, this would make performance *transient* as the individual difference only influenced performance during early stages of learning (Keehner et al., 2006). Confirmation that a specific individual difference causes *transient* performance outcomes, within a field, is useful information for assessment as it confirms that such a difference should not be used for initial screening (Keehner et al., 2006). If performance is *enduring, then an* individual difference would still influence performance even after practice (Keehner et al., 2006). A discovery of an individual difference that causes an *enduring* performance outcome for a skill is also informative as it could be used to predict possible future performance by using pre-screening and one-time testing (Keehner et al., 2006).

Furthermore, research into individual differences and innate ability is paramount. Especially as many studies have found that for surgical training, although most participants will respond to

training and improve. There is still a small group of individuals that will never improve even with repetitive practice (Alvand, Auplish, Khan, Gill, & Rees, 2011; Grantcharov & Funch-Jensen, 2009; Louridas et al., 2017; MacMillan & Cuschieri, 1999; Schijven & Jakimowicz, 2004) . According to MacMillan and Cuschieri (1999) this incapacity to improve can be explained by innate ability which is the "level of aptitude (qualities that an individual brings to a task by virtue of his/her innate genetically determined ability)". For example, Schijven and Jakimowicz (2004) put individuals that do not improve with practice into two groups. One group that has such a strong innate ability that they start off strong and don't need practice to improve; and another group that has such a low innate ability they cannot build the psychomotor skills necessary to perform laparoscopy surgery. Finding individual differences that may be able to distinguish those with an innate ability and those without, would be crucial step to provide more accurate assessment, and vital for creating prediction tools and algorithms that can help both recognise and guide those that have inadequate technical skills (Grantcharov & Funch-Jensen, 2009). However, to create these prediction tools for the future, more information is needed. The main limitation to these past studies is that they focused on group level performance, and while a phenomenon may occur on a group level, this may not automatically imply the same phenomenon will be represented at an individual level. Therefore, in order to actually determine if there are individuals that will never improve, research into individual differences must also be explored where the individual level is taken into account. This should be done either through multilevel modelling that can account for both within-person and between-person variation, and/or exploration into each participants individual learning curve (Hoffman, 2007; Schmettow, 2018).

4. Individual Differences

Skill acquisition can vary from person to person. As people are different, when learning a task one person may have a starting advantage, be faster, more accurate, or acquire proficiency faster compared to another person. Pinpointing which individual differences may be responsible for variations in performance can be useful for assessment and training. Past research pertaining to

gender and initial performance are outlined below, and the influence these two individual differences have shown to have on performance.

4.1 Gender

Past research in the laparoscopy domain has indicated that male participants have an advantage in regard to both speed and visuo-spatial abilities (Donnon et al., 2005; Grantcharov, Bardram, Funch-Jensen, & Rosenberg, 2003; Thorson et al., 2011). However, gender differences in accuracy and efficiency have mixed findings with some research showing a male advantage and other research showing no gender differences. The study by Thorson et al. (2011) on medical students with no laparoscopy experience found that, when exposed to a simulator, female participants created more errors than male participants. Additionally, Groenier et al. (2015) found that male participants had more efficient movements compared to females. On the other hand, research by Grantcharov et al. (2003) found that for tasks on a virtual reality laparoscopy simulator errors between gender groups were not different. To conclude, although a male advantage has typically been found especially for speed, the influence that gender may have on accuracy performance is unclear.

The current study aims to expand our understanding of gender differences by following the recommendations by the Institute of Medicine (IOM) (2012) to undertake gender-based research. The IOM emphasised that currently, females are underrepresented in scientific research, and correct statistical reporting of gender differences is lacking.

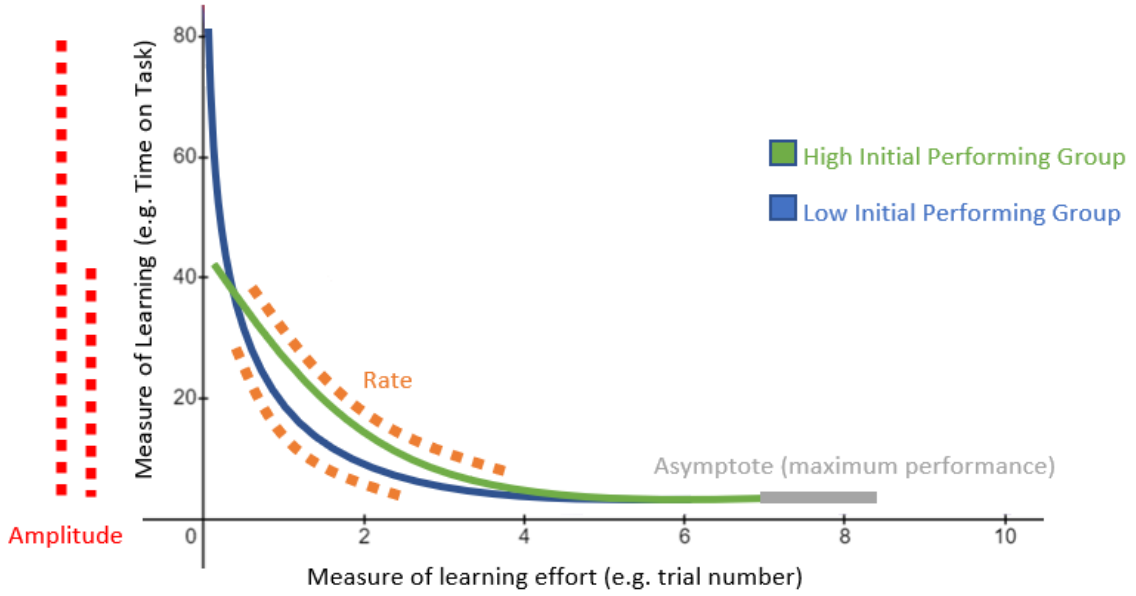
4.2 Initial Performance

Past research has found that participants with the same level of initial proficiency in a task may differ in later performance Bahrick et al., 1993). Hence, initial proficiency does not determine later performance for an acquired skill. As shown in research by Adams (1957), this outcome can be explained by three learning curve parameters: *amplitude* which shows the performance range from initial performance to maximum performance; r*ate* which is determined by how fast a person learns; and *asymptote* whereby an individual's maximum performance has been achieved. In the study by

Adams (1957), as shown in Figure 2, the group with poor initial performance had a high *amplitude*. However, as they displayed a high learning *rate* it led to an *asymptote* with optimal end performance. The high initial performance group had a low *amplitude*, nonetheless, as it was coupled with a low learning *rate.* This counteracted their initial starting advantage, and led to end performance that was optimal. Most importantly, they found that the members of the high initial performing group displayed similar *asymptotes* to participants in the low initial performing group.

Figure 2

Learning Curves of the High Initial Performing Group vs. Low initial Performing Group



Note. The figure displays the findings by Adams (1957), where the *low initial performing group* had a high amplitude and a high learning rate; while the *high initial performing group* had a low amplitude, but a low learning rate. The differences in amplitude and rate produce end performance outcomes (asymptotes) which were similar for members of the low initial performing group and the high initial performing group.

This indicates that for learning, one parameter alone is not enough to obtain an overall picture of skill acquisition and performance, and instead each individual parameter must be interpreted in light of the other parameters. Therefore, selection and assessment procedures based on scores only during initial intake may not represent the whole picture. In fact, such selection procedures may be detrimental, as candidates who with practice hold the potential to obtain optimal scores may be overlooked.

Learning curve data is typically complex as it is both *longitudinal* – with measurements occurring at different time points, as well as *non-linear* – as a participant does not produce equal improvement from one session to the next. Complexity further increases if there is also a focus on individual differences at a participant level, rather than just at a group level where information can be lost during aggregation (Bürkner, 2017; Schmettow, 2018). Appropriately, capturing these complexities in a model would be a struggle for standard statistical approaches to compute, and rather a specific approach is necessary (Bürkner, 2017).

5. Bayesian Statistics

The current study is distinct as instead of using previous Frequentist traditional approaches to data analysis, a Bayesian approach was used. A Bayesian approach is advantageous as it can allow complex model which can handle non-linear data, while a Frequentist approach is better suited to linear modelling (Bürkner, 2017). The Bayesian approach also has the advantage that it is better suited to deal with smaller sample sizes (Institute of Medicine (IOM), 2012; Zhang, Hamagami, Lijuan Wang, Nesselroade, & Grimm, 2007).

Although both types of data analysis can be used for multilevel modelling, there are two main distinctions (Bayarri & Berger, 2004). Firstly, how probabilities are viewed are different in each analysis (Schmettow, 2018). In traditional statistics, the aim is to reach a *p-value* at a confidence interval of 95% (Bayarri & Berger, 2004). The goal is hypothesis testing whereby one either accepts or rejects a hypothesis. Although this may seem intuitive on the surface, the underlying meaning is not straight forward. From a Frequentist approach, the idea of probability is based on confidence intervals (Bayarri & Berger, 2004; Schmettow, 2018). For example, if 100 hypothetical experiments were to take place, then at least 95 of these experiments would include the true value. Bayesian statistics is more intuitive and based on credibility intervals. The rationale is not to undertake hypothesis testing but make inferences and parameter estimates where the researcher can be 95% certain that the true value or population mean is within an interval range (Schmettow, 2018). Secondly, a main distinction is the use of prior knowledge (Smith & Gelfand, 1992). From a

Frequentist view, prior knowledge is of no importance and all data is random, while for a Bayesian approach modelling can be based on prior knowledge that was learnt from the data either currently or in the past (Smith & Gelfand, 1992). For example, a Bayesian model will output a posterior distribution, which is essentially a prior distribution that has been changed with the addition of data (Phillips, 1975). This posterior distribution can be used to incorporate learnt knowledge into new models, by using the information from the posterior as priors for future models – which hold the potential to be used for forecasting and prediction (Phillips, 1975).

As previously noted, Bayesian statistics – although more intuitive – does appear to have certain limitations. Firstly, it increases the researcher's degrees of freedom (Simmons, Nelson, & Simonsohn, 2011). Therefore, as the researcher can give more input, this increase in choice could bring about unethical practices, as it may provide results that match what the researcher was aiming to find (Simmons et al., 2011). For example, a researcher has the freedom to choose what type of prior distribution they may want to use for the data. This is further made difficult as there is no objective principle in place that helps to decide on picking a non-informative prior as a distribution, or how to pick an informative prior (Gelman, 2008). The article by Gelman (2008) points out other limitations such as prior and posterior distributions are based on subjective knowledge instead of objective facts. Furthermore, these subjective prior distributions may not transfer well to each situation. In addition, there is a high reliance on assumptions which can lead to biased results. Lastly, the addition of multilevel modelling can also complicate the data and lead to even more assumptions.

Due to these limitations, *model fit* is crucial in Bayesian statistics to avoid making incorrect assumptions. For example, analysis can be done to determine if the assumed distribution chosen by the researcher is in fact the appropriate choice to represent the raw data (Bürkner, 2017).

The Current Study

The primary aim of the study was to model learning curves to explore the influence that gender and initial performance had on skill acquisition, when using a laparoscopic simulator. The current

study was a secondary analysis from pre-existing data which was previously used in two published studies by Groenier et al. (2014) and Groenier et al. (2015). The past studies were mainly concerned with cognitive ability and gender (Groenier et al., 2015, 2014). While the current study did not explore cognitive ability, its aim was to focus on how gender and initial performance influence skill acquisition on LapSim.

The secondary aim was to use a Bayesian approach. The two past studies used multilevel modelling but from a Frequentist approach. Nevertheless, as learning curve data is longitudinal and non-linear, it is proposed that the Bayesian approach should be more appropriate for fitting such data. While a Frequentist approach uses p-values to determine the presence of an effect. A Bayesian approach makes estimations from the outputted fixed effect parameters obtained from the posterior distribution. In combination with credibility intervals, it can be determined if an effect was present or not.

The overall motivation of the study is to provide a greater understanding of individual differences, which is important for assessment and determining if there is a need for individualized training programs.

Using a Bayesian approach and estimation of the posterior distributions, the following research questions were investigated:

RQ1: Do individual differences, like gender and initial performance, produce performance outcomes which are *transient or enduring?*

RQ2: To what extent does gender influence the learning curve for duration and accuracy performance, as skill acquisition occurred on a laparoscopic simulator; and additionally, how robust this influence is based on model criticism and model fit?

RQ3: To what extent does initial performance of duration or damage count, influence the learning curve of the respective performance measure, as skill acquisition occurred on a laparoscopic simulator; and additionally, how robust this influence is based on model criticism and model fit?

Method

Procedure

A previous description of the procedure section can be found in the Groenier et al. (2014) and Groenier et al. (2015) papers. The studies were part of a longitudinal study with repeated measures. Participants did weekly 30-minute training sessions, during these sessions participants practiced basic LapSim tasks for the allocated time. The number of observations/trials varied for each participant as faster participants completed more tasks, and slower participants less tasks.

As the training sessions were a proficiency-based program, the number of sessions also varied between participants, as after passing examinations participants were no longer required to continue the study. Analysis only included data of sessions up till the 5th session, while the previous papers also used session 6; this change was done as session 6 had missing data due to some participants finishing the training program.

In the current study, *duration* and *damage count* were the two variables that were outputted for analysis. Both variables were used to measure performance as participants did the medium difficulty level for two tasks: *grasping* and *instrument navigation.*

Participants

The current study used 75 participants; this is less than what was used in the previous studies as unexpected circumstances meant that the data recorded for the previous studies could not be utilised. Eight participants did not have their gender recorded and were excluded from the analysis. Of the remaining 67 participants, 38 were female, and 29 were male. The average age was 22.66 years (*min.* = 20, *max.* = 26, SD = 1.32).

Apparatus - LapSim

Two LapSim simulators had the same setup, with participants randomly assigned to either simulator. LapSim v.3.0.10 was used, which was produced by the company Surgical Science. The set up included Immersions VLI hardware and a 19-inch computer monitor that displayed the virtual

surgical environment. Furthermore, feedback from the input instruments was mirrored onto a monitor.

LapSim has been validated as a generally sound assessment tool. A study by Van Dongen, Tournoij, Van Der Zee, Schijven, and Broeders (2007) showed that LapSim can be used as a successful assessment tool for evaluation of laproscopy performance as it can differentiate distinct groups. In the study, it is reported that LapSim could differentiate between novices, residences in training with some laparoscopy experience, and experienced laparoscopy surgeons.

Performance Variables

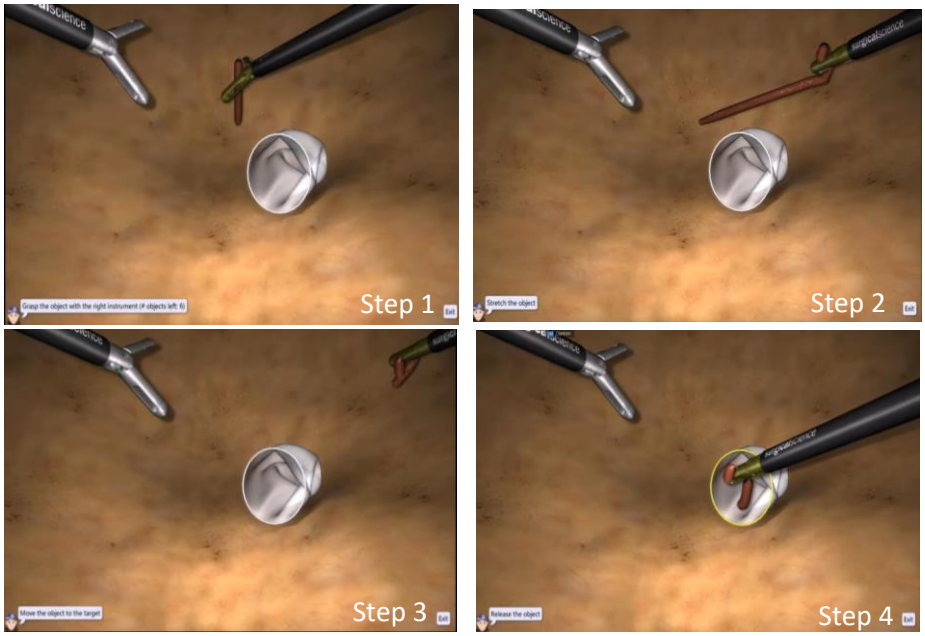
Duration and damage count were the two variables used to measure performance. *Duration* was a combined value which was the average of *right-hand time* and *left-hand time,* with both measures being in seconds. Damage and accuracy were assessed using *damage count,* which is the number of errors a participant made in each trial. For both performance variables a low value indicated better performance (shorter duration and fewer errors) compared to a higher value.

Tasks

Grasping. For this task, there is an object that is connected to the tissue wall. The simulator asks the participant to grasp the object. Once they grasp the object, they are then supposed to stretch it until it becomes disconnected from the tissue wall. The participant then moves the object into an endoscopic bag. This process is repeated, but the object will appear in different places on the tissue wall, and the instructions will tell the participant to alternate the hand they are using. Refer to Figure 3 that shows images of the task.

Figure 3

Steps for Grasping Task

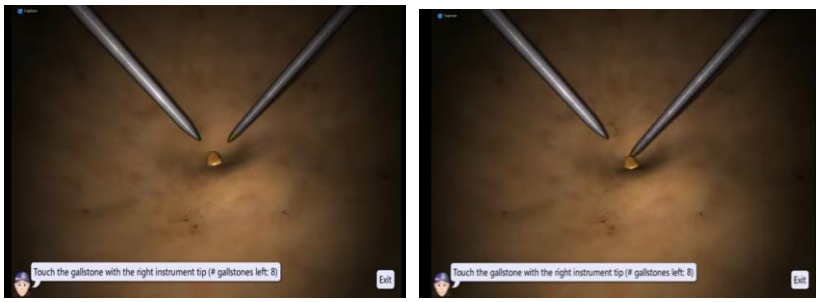


Note. The images were taken from a *YouTube* video uploaded by Surgical Science (2012).

Instrument Navigation. For this task, a gallstone will appear. The task has a time limit and the goal is to use the instrument tip to touch the gallstone before it disappears. This process is repeated but the gallstones will appear in different places on the tissue wall, and the instructions will tell the participant to alternate the hand they are using. Feedback in the form of a yellow highlight also helps the participant determine which hand to use. The instructions also count down how many gallstones are left. Refer to Figure 4 that shows images of the task.

Figure 4

Steps for Instrument Navigation Task



Note. The images were taken from a *YouTube* video uploaded by Surgical Science (2012b).

Statistical Analysis

All statistical analysis was done using R version 3.6.1 combined with R studio, with a *tidyverse* setup. Appendix A gives an overview of what R libraries were used, what they were used for, as well as main functions used from those libraries.

Participant Exclusion

An outlier removal procedure was put into place. The criteria for removal of a trial was any trial that had an extreme value for any performance measure. An extreme value would be a result that was more than 3 standard deviations (SD). Furthermore, if an extreme value was displayed for three consecutive sessions, then all trials for that participant were to be removed. No participants or trials were removed based on the outlier removal procedure.

Respondents who did not provide gender demographics were excluded; this included 8 participants, leaving 67 participants. Of these participants, three more were excluded from all initial performance models as they did not have recorded data in session 1.

High vs. Low Initial Performers

Quartile grouping was used to create groups of high and low initial performance groups. The grasping tasks and instrument navigation tasks from *session 1* were taken together to obtain initial performance. The quartile groups were made specific to each performance measure (e.g. duration, damage count). For each individual, an average of the specific performance variable was made using all their trials in the first session. For example, when looking at duration, the quartiles were made by having an average duration score for each participant (for session 1) and then ordering these from best performance indicated by a participant having a fast average with a low number of seconds, to worse performance where a participant had a slow average and a high number of seconds. From this ordering, 4 (mostly) equal-sized participant groups were created. The group division was produced using the r function called *"quantile"*. In the case that the groups were uneven, this function determined how the groups were divided into 4 subgroups with a mostly equal number of participants. The 2nd and 3rd quartiles in the model were filtered out as the current study was only

primarily interested in the high initial performing group and low initial performing group (1st and 4th quartile group, respectively). For the damage count performance variable, the same approach was taken. The high initial performing group was the 1st quartile group and included participants that had a low average damage count in the first session, and therefore had high accuracy. The low initial performing group was the 4th quartile group, that had participants who obtained a high average damage count in the first session, and therefore had low accuracy. Three participants were not included in the analysis as they did not have data in session 1.

Data Exploration

The population level performance outcomes, for all the groups (male group, female group, low initial performance groups, high initial performance groups) are shown in Appendix B. This includes the median, mean, standard deviation, minimum value and maximum value, of the damage count and duration scores for each session.

Checking Assumptions and Data Understanding. Three aspects were explored to either check assumptions or obtain a greater understanding of the data, and can be found in Appendix C. Firstly, histograms were used to determine if the performance variables were normally distributed. A common assumption for many frequentist and parametric tests (such as ANOVA and t-tests) is that performance data is normally distributed (Schmettow, 2018). In certain instances, when this assumption is violated, many researchers will continue to use these parametric tests even when non-parametric tests would better handle such a violation (Schmettow, 2018). Therefore, data not normally distributed would not fit a linear model well. A method to tell if normality has been violated is to use histograms which can show if the raw data is normally distributed or skewed.

Secondly, over-dispersion was checked. Overdispersion occurs when there are data values that are not as frequent are seen at the end of the tail of a distribution. The main reason for exploring dispersion is that it will influence how the data is modelled and if there is correct model fit. This is especially crucial for the damage count data that uses a Poisson distribution. If overdispersion

is found, then the model needs to be modified by adding an *observation level random effect* (Schmettow, 2018). This addition will change the model by making the intercept of the damage count dependent upon each trial/observation, and consequently, trials near the tail of the distribution will be included in the analysis (Schmettow, 2018). Overdispersion can be seen visibly in a histogram, however as modelling the damage count data correctly is reliant on recognizing overdispersion, an additional overdispersion test was performed for this data. The test was done using the R Library *AER*. The test first used a Poisson model to fit the data; this model was then tested using the function *dispersiontest.*

Lastly, violin plots were used to detect if there is variation between the gender groups, as well as variation between the initial performance groups. It also determined if this variation was affected by the type of task (either grasping or instrument navigation).

Graphical Representation of Learning Curves

The current study used raw data to make a graphical representation of the learning curves. The x-axis was the number of sessions, while the y-axis used either the median or mean of a performance variable. The study mostly used the median as the central tendency measure, especially for exploration, as it is more accurate with data that is not normally distributed (Schmettow, 2018)**.** The mean was only used for the duration models as this was the outputted default.

Multilevel Exploration

It is recommended by Schmettow (2018) to perform multilevel exploration. The analysis for this exploration can be found in Appendix D. The first step was to overlap each individual learning curve, known as the participant effect, with that of the population group level learning curve, known as the population-level effect. From visual inspection, it can be used to infer if overall differences amongst participants are transient or enduring depending on variation in performance. If the participant effects are more varied in the beginning and become more converging with more

practice, it can be assumed that participants overall differences became less pronounced with practice. However, this does not give an indication regarding which specific individual differences are transient or enduring.

The second step was to explore individual learning curve graphs for each participant and establish if aspects of the group level learning curve also adhere at an individual level. If there are individuals that do not present the same group effect, this may cause inconsistencies when data is analysed at the group level. Having too many outliers may negatively influence results. The overall aim of this part is to determine using visual inspection how many participants do not follow the population effect.

Analysis was also done to determine the noise created by the individual. Results regarding overall noise in duration and damage count can be found in Appendix E (Table E1, and Table E2). The library *rstanarm* was used with the function *stan_glmer* to make a reference group (intercept) that was dependent on each individual. The function *coef* was used to obtain a sigma measure, with a higher sigma value indicating more individual noise. Table E3, and Table E4, has the *ranef* output which gives the results based on the predicted posterior distribution for each participant, which provided an indication of which participants produced the noise.

Multilevel Modelling

The multilevel models in the current study were based on recommendations from the book, *New Statistics for the Design Researcher* by Martin Schmettow (2018). The reasoning for using a specific model and its chosen prior distribution follows the logic the author set out. The two R packages used were *rstan* and *brms* that run on an R interface, but implement a probabilistic programming language known as *Stan* as a backbone to run the models (Bürkner, 2017; Stan Development Team, 2020).

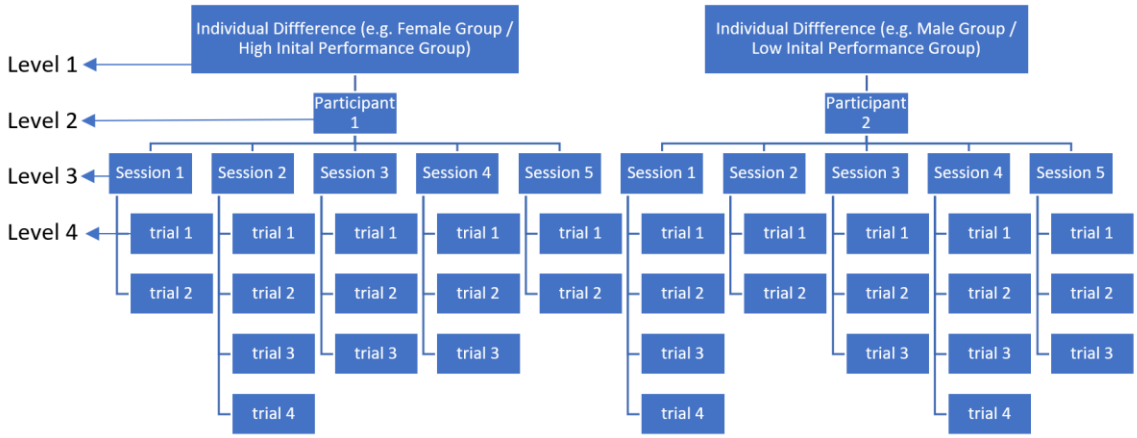
Four multilevel models were utilized in the current study, half the models fitted duration data and the other half fitted damage count data. Learning curve data is typically non-linear and

longitudinal, and records skill acquisition as it occurs over time. This leads to data that has many levels. Firstly, to accurately model learning it is necessary to take account of all these different levels (Zyphur, Kaplan, Islam, Barsky, & Franklin, 2008). Multilevel modelling is a great approach for modelling this type of learning curve data, as it has the ability to incorporate levels where "individual data is nested within groups" (Zyphur et al., 2008). The multilevel models in the study had four levels, see Figure 5. Two models used gender groups as the top level, and the other two models used initial performance grouping as a level. This levelling allowed exploration into these specific individual differences and their influence on skill acquisition. The participants/individuals were placed on the next level and were categorized based on the groups they belonged to. This level accounts for both between and within person variation (Hoffman, 2007). All the models then had the session number placed as a lower level. This allows the learning curve to be approximated by a statistical model, and therefore exploration of skill acquisition and how the different groups progressed with practice can be interpreted. The lowest level consisted of each trial and repetition an individual did. This level is important as the trials are often highly repetitive and done multiple times by the individual. As an individual repeats the same task we get non-independence, which is the concept that future trials are influenced by past trials (Zyphur et al., 2008). However, most standard statistical approaches assume that trials are independent and are not influenced by each other (Zyphur et al., 2008). One of the main advantages of multilevel modelling is that assumes non-independence and therefore more suitable for data that incorporates learning.

The mean performance score of all the repetitions for a given session was utilised for the duration models, while the mean function was used for the damage count models. However, as the sessions did not have a fixed number of trials, the number of repetitions could vary. Consequently, if a participant was faster, they may have completed more repetitions and had more observations for a given session compared to a slower participant.

Figure 5

Conceptual Representation of the Levels in the Multilevel Models



Note. The top-level is the individual differences, gender was split into a female group and male group, while initial performance was split into a high initial performing group and a low initial performing group. The next level is the participant, which is followed by the session level, and then proceeded by a level that contains the number of trials/observations which was not fixed and varied between the participants. The figure above is conceptual outline of the multilevel models and has been simplified for clarity, therefore it does not show all the components of the models (e.g. all participants have not been added).

Duration Multilevel Modelling. For duration data, the R library *brms* was used with the function *brm* to create a multilevel model with a prior distribution (Bürkner, 2017). The prior distribution used was an exponentially modified Gaussian distribution, known as an ExGaussian distribution. According to Schmettow (2008), an ExGuassian distribution has three parameters making it ideal for time-related data as it can account for skewed distributions that also have a large dispersion. The three parameters allow location and dispersion to vary independently. This is different compared to Poisson, Binomal and Exponential distributions where dispersion and location are dependent upon each other, or even the same value (Schmettow, 2018). To interpret the model, the output function *fixef* was used. This gave the fixed effect parameters of the posterior distribution, by providing the mean regression coefficient. This can then be used to draw contrasts between the different levels. A summative analysis took place whereby time is added or subtracted to the reference group (intercept) to determine group level differences. Therefore, a negative value indicated performance improved as there was an increase in speed and participants took less time to

complete the tasks; a positive value indicated that performance became worse as there was a decrease in speed and participants took longer to complete the tasks.

Damage Count Multilevel Modelling. For damage count data, the R library *rstanarm* was used with the function *stan_glmer* to create a multilevel model with a prior Poisson distribution (Stan Development Team, 2020). According to Schmettow (2008), this distribution is advantageous as count data cannot have negative values but is instead bounded at zero. This more closely resembles the real-world data measured. As previously mentioned above (see "Data Exploration" *observation level random effect*), if the damage count data is overdispersed, then an *observation level random effect* would need to be added to the model (Schmettow, 2008).

To interpret the model, the output function *fixef* (fixed effect parameters) was used with an *exponential mean function*. This is because the logarithmic scale cannot be interpreted directly and the exponential mean function enabled expected fixed effect values to be obtained from the model's posterior distribution (Schmettow, 2018). A multiplicative analysis took place, to determine group-level differences which are based on rates, and percentages. The output of the model is interpreted by the value being either above or below 1. If the value was above 1, performance became worse as there was an increase in errors made; a value below 1 indicated that performance improved as there was a reduction in errors.

Model Criticism and Model Fit. The Bayesian models were checked for model criticism and model fit. Four main aspects were analysed and can be found in Appendix F. Firstly, grouping and dispersion of the groups were checked by using the predictions that were outputted by the model. Secondly, residual (standard deviation) analysis was conducted to check for variation between the groups, as differences can create inaccurate predictions. Residuals are made using the observed measure and the predicted measure and can be calculated as taking the observed score minus the predicted score. This indicates variability of the sample from the population. Overall, the larger the residuals, the less the model predictions can be trusted. Thirdly, analysis for the

predictive power of the model took place, whereby the model with an individual difference as a level is compared to another model that only had sessions as a level. This was done to determine if an individual difference created different groupings with different predictions compared to if no level had been used. If predictive ability is found it indicates that the credibility intervals can determine that there are differences between the groups, as differences in proportions between the groups were large enough. Therefore, the model would be useful for decision making as the outputted posterior distribution could be used as a prior distribution for a predictive model.

Lastly, the model fit of the chosen distribution was conducted. The data was fitted with many distributions to determine if the distribution used had the best fit for the raw data compared to other possible distributions. In Bayesian statistics, a good fit is necessary to increase the likelihood of obtaining a valid statistical model, where credibility intervals are more exact.

For model fit, the duration and damage count models had different chosen distributions used for comparison. For the duration models the distributions chosen were ExGaussian, Gaussian and Gamma. All these distributions were added as prior distributions into a multilevel model, using the R library brms and the functions: brm, post_pred, and posterior. Other functions used were GGplot for making figures which used the R library tidyverse. For the damage count models the distributions chosen were Negative Binomial and Binomial. Q-Q Plots were made using the r library MASS and vcb with the function displot whereby the distribution type could be chosen. A Q-Q plot is created by plotting observed frequency over the fitted frequency of the chosen distributions. A model had a good fit if the observed data matched that of the theoretical distribution.

Results

Data Exploration

The relevant figures for this section are in Appendix C. Firstly, the histograms showed that the duration data had distributions with a bimodal peak (Figures C1 and C2), while the damage count data was right-tailed with a unimodal peak (Figures C3 and C4). As all performance variables were

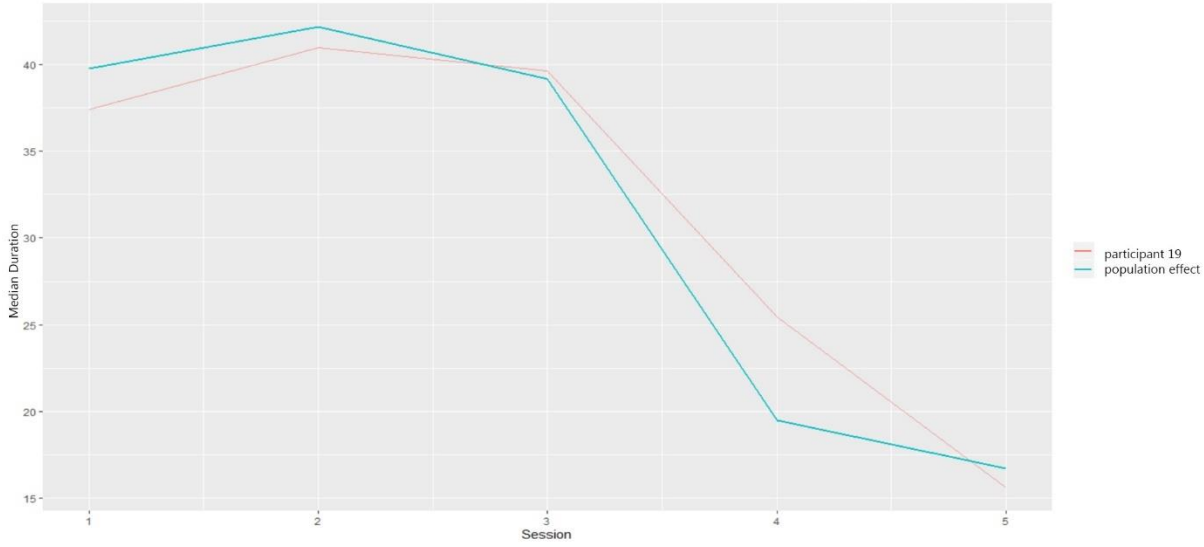
not normally distributed, they violate the assumption of normality, making the data inappropriate for parametric tests and linear modelling. Secondly, all the data indicated overdispersion. Thirdly, the violin plots showed some differences in variation between the groups (Figures C5, C6, C7 and C8). However, the variation was not considerable enough to affect the type of distributions (e.g. bimodal, right-tailed) and consequently any differences should not have a considerable influence on the results.

Multilevel Modelling

Through multilevel exploration (see Appendix D), from visual inspection it appeared that most individuals have learning curves that match closely to the general population effect for both duration and damage count. Figure 6 shows an example of a participant in the study and how their individual learning curve closely matches that at the population level. Therefore, multilevel non-linear modelling is appropriate to utilize for gender and initial performance analysis as it is assumed that even at an individual level the data is non-linear.

Figure 6

Participant 19's Median Duration Learning Curve and the Group Level's Median Duration Learning Curve (Population Effect)



Note. The figure shows participant 19's median duration scores for the first 5 sessions. From visual analysis this individual's learning curve is similar to the learning curve at the population level. Therefore, non-linear curves were seen at both an individual level as well as at the population level.

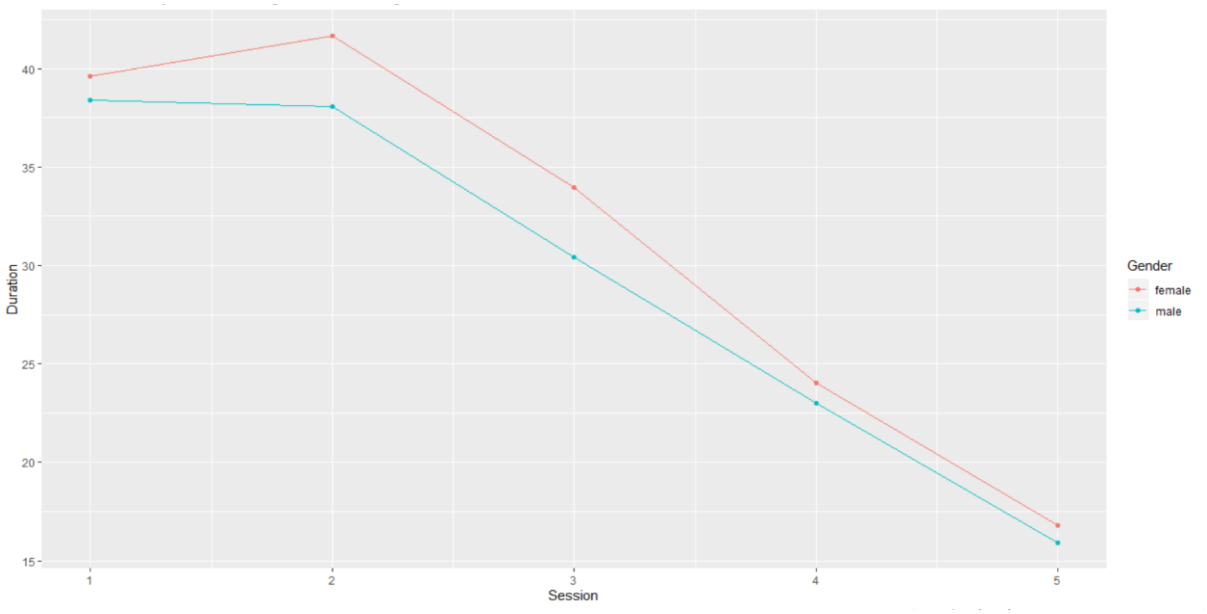
Appendix G holds the raw *fixef* (fixed effect parameter) outputs which were outputted from the posterior distribution created by the multilevel Bayesian models; as well as the calculations needed to obtain understandable values.

Gender Groups

Duration. Visual inspection of the raw learning curve showed that the female group mean duration was consistently higher than the mean of the male group, indicating they were consistently slower (Figure 7). The multilevel model, from Table 1, cannot confirm that the male group held a starting advantage over females. It can however confirm that the female group became faster from session 3 and onwards, thereby gaining an advantage when practising. The model cannot confirm that male participants were faster than the female group for any given session.

Figure 7

Progression of Sessions, showing Mean Duration for Male vs. Female Groups



Note. A higher duration score indicates that the group was slower. In this respect, the red line, which is the female group mean shows that they were consistently slower than male group mean (blue line).

Table 1

Duration Model Overview for Gender Groups

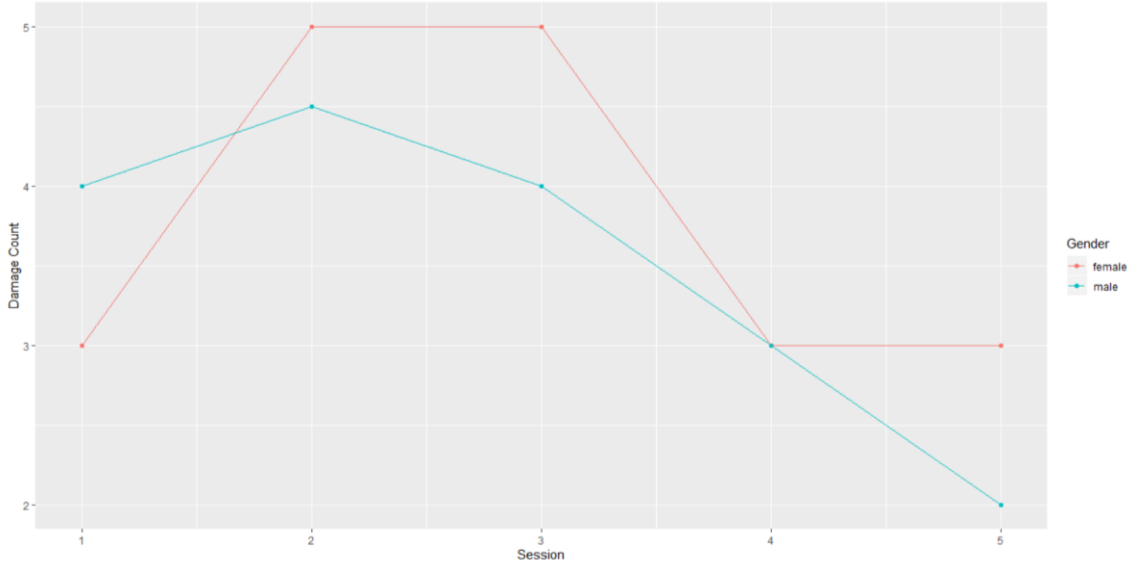
Session	Mean Time (in seconds)		Credibility intervals 95%		Credibility interval Assumptions	
	Female Group	Male Group	Female Group	Male Group	Female Group Duration compared to Session 1	Male Group Duration compared to Female Group Duration
1	39.63	38.36	[37.84, 41.33]	[-4.17, 1.60]	N/A	N/A
2	41.47	38.16	[-0.51, 4.17]	[-5.71, 1.61]	Not known	Not Known
3	33.58	30.63	[-8.26, -3.78]	[-5.21, 1.75]	Faster	Not Known
4	23.46	23.00	[-18.31, -13.79]	[-2.82, 4.31]	Faster	Not Known
5	16.81	15.92	[-25.20, -20.40]	[-3.51, 4.19]	Faster	Not Known

Note. Credibility assumptions were made based on if the credibility interval was negative or positive. If both the upper and lower bound were negative then the group were faster, if the lower bound was negative and the upper bound positive then it is unknown if the group was faster or slower, if the upper and lower bound are positive then the group was slower. This data was made using the fixed effect output of the posterior distribution, and can be found on Table G1, and calculations are on Table G2.

Damage Count. Visual inspection of the raw learning curve (Figure 8) showed that male participants on a group level had a higher number of median errors than female participants in the first session. However, as the sessions progressed, male participants improved and by session 5, they had a lower median damage count compared to female participants. The multilevel model, from Table 2, can confirm that the female group had a starting advantage in their first session over the male group in terms of damage count. It can also confirm that the female group showed no progress in the beginning sessions. Nonetheless, with practice, the model can also confirm the group improved their accuracy. At session 5, it is certain that the female group managed to reduce their rate of damage count.

Figure 8

Damage Count Session Progression, showing Median Damage Count for Male vs. Female Groups



Note. A higher duration score indicates that more errors were made. For session 1, the male participants on a group level made more median errors than female participants. However, as sessions progressed, male participants improved, and by session 5, they had a lower median damage count compared to the female group.

Table 2

Damage Count Model Overview for Gender Groups

Session	Damage Count (using Rate)		Credibility intervals 95%		Credibility interval Assumptions	
	Female Group compared to Session 1	Male Group compared to Female Group at Session 1	Female Group	Male Group	Female Group Damage Count compared to Session 1	Male Group Damage Count compared to Female Group Damage Count
1		3.4x more errors	N/A	[0.72, 1.33]	N/A	N/A
2	1.53	0.90x less errors	[1.24, 1.86]	[0.65, 1.25]	More Errors	Not known
3	1.37	0.84x less errors	[1.07, 1.76]	[0.56, 1.24]	More Errors	Not known
4	0.81	0.98x less errors	[0.62, 1.07]	[0.66, 1.49]	Not Known	Not known
5	0.73	0.8x less errors	[0.58, 0.91]	[0.57, 1.17]	Less Errors	Not known

Note. Credibility assumptions were made based on if the credibility interval was greater or lower than 1. If both the upper and lower bound were below 1 then the group improved and had less errors. If the lower bound was below 1 and the upper bound above 1 then it is unknown if the group had a performance increase or decrease. If the upper and lower bound are more than 1 then the group had an accuracy performance decline and made more damage errors. This data was made using the fixed effect output of the posterior distribution, and can be found on Table G3.

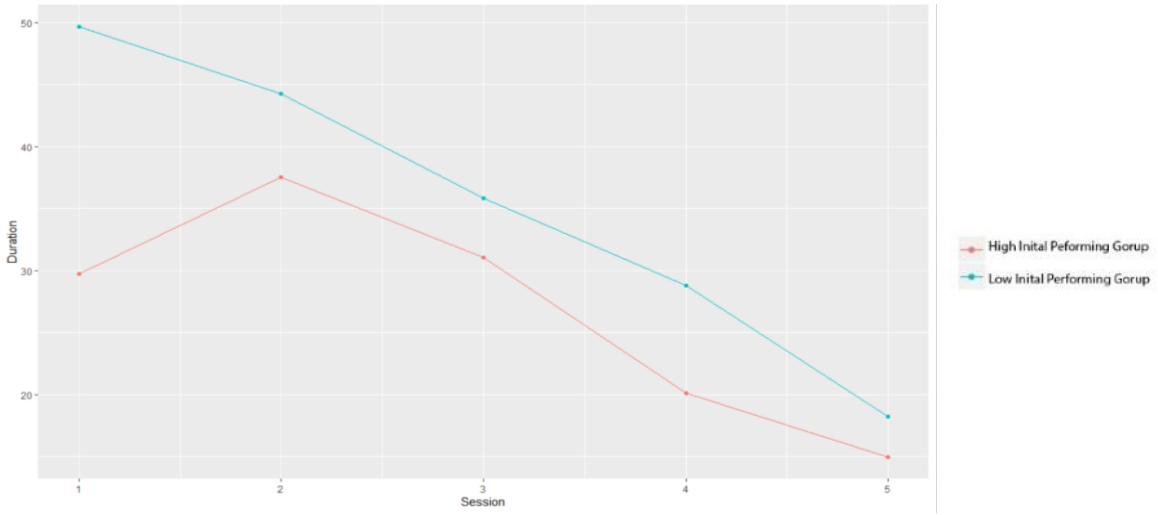
Initial Performance Groups

Duration. Visual inspection of the raw learning curve showed that the high initial performing group consistently had lower mean duration scores, indicating they were faster than the low initial performing group (Figure 9). The multilevel model, from Table 3, can confirm that the high initial performing group held an advantage in the beginning. It can also confirm that this group also had an increase in time and were slower for session 2, however by session 4, they showed a definite reduction in duration and got faster.

The model can also confirm that the low initial performing group improved at such a substantial rate that, at the end, their starting disadvantage no longer influenced the results. For session 2, 3 and 4, it is apparent that the low initial performance group showed slower duration scores compared to the high initial performing group. However, by session 5, the credibility intervals were too wide and overlapping, and it is doubtful whether this advantage for the high initial performing group remained.

Figure 9

Duration Session Progression, showing Mean Duration for High Initial Performers vs. Low Initial Performers



Note. A higher duration score indicates that the group was slower. In this respect, the red line, which is the mean of the high initial performing group was consistently faster than the mean of the low initial performing group (blue line).

Table 3

Duration Model Overview for Initial Performance Groups

Session	Mean Time (in seconds)		Credibility intervals 95%		Credibility interval Assumptions	
	High Initial Performing Group	Low Initial Performing Group	High Initial Performing Group	Low Initial Performing Group	High Initial Group Duration compared to Session 1	Low Initial Group Duration compared to High Initial Group Duration
1	29.76	49.68	[26.98, 32.70]	[15.93, 23.74]	N/A	N/A
2	37.55	44.29	[4.04, 11.25]	[-18.07, -7.96]	Slower	Faster
3	31.06	35.74	[-2.28, 4.87]	[-20.14, -10.34]	Not Known	Faster
4	20.06	28.80	[-13.20, -6.25]	[-16.07, -6.38]	Faster	Faster
5	15.01	18.24	[-18.51, -11.45]	[-21.66, -11.45]	Faster	Faster

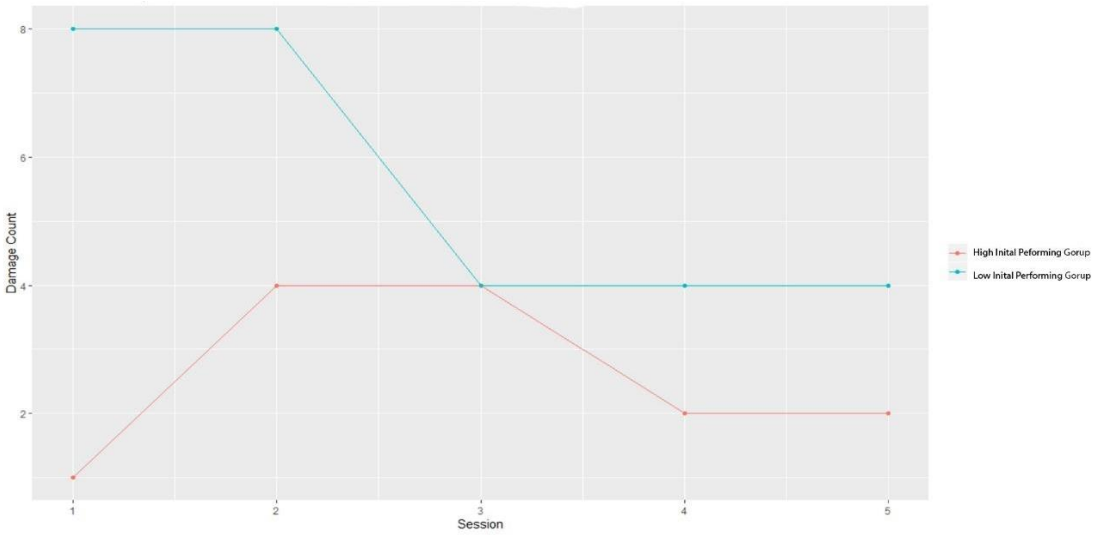
Note. Credibility assumptions were made based on if the credibility interval was negative or positive. If both the upper and lower bound were negative then the group were faster, if the lower bound was negative and the upper bound positive then it is unknown if the group was faster or slower, if the upper and lower bound are positive then the group was slower. This data was made using the fixed effect output of the posterior distribution, and be found on Table G4, and calculations are on Table G5.

Damage Count. Visual inspection of the raw learning curve showed the high initial performing group consistently produced fewer errors than the low initial performing group, except for session 3, where both groups at the population level have the same median number of errors (Figure 10). The multilevel model, from Table 4, can confirm that the high initial performing group had a starting advantage. It can also confirm that this group lost this starting advantage as they had an increase in damage count at a group level as sessions progressed compared to a decrease.

For the low initial performance group – the model can confirm that at a group level, as sessions progressed, they were able to improve damage count for each session at a faster rate compared to that of the high initial performing group.

Figure 10

Progression of Sessions, showing Median Damage Count for High Initial Performers vs. Low Initial Performers



*Note***.** A higher damage count score indicates that the group made more errors and were less accurate. Hence, the red line which is the high initial performing group shows that they consistently produced fewer median errors than the low initial performing group (blue line), except for session 3 where both groups at the population level had the same number of median errors.

Table 4

Damage Count Overview for Initial Performing Groups

Session	Damage Count (using Rate)		Credibility intervals 95%		Credibility interval Assumptions	
	High Initial Performing Group compared to Session 1	Low Initial Performing Group compared to High Initial Performing Group at Session 1	High Initial Performing Group	Low Initial Performing Group	High Initial Group Damage Count compared to Session 1	Low Initial Group Damage Count compared to High Initial Group Damage Count
1		1.38x more errors	N/A	[1.01, 1.85]	N/A	N/A
2	3.18	0.27x less errors	[2.20, 4.68]	[0.17, 0.44]	More Errors	Less Errors
3	2.90	0.22x less errors	[1.91, 4.46]	[0.13, 0.38]	More Errors	Less Errors
4	1.56	0.31x less errors	[2.02, 2.39]	[0.18, 0.55]	More Errors	Less Errors
5	1.44	0.28x less errors	[0.96, 2.14]	[0.17, 0.47]	Not Known	Less Errors

Note. Credibility assumptions were made based on if the credibility interval was greater or lower than 1. If both the upper and lower bound were below 1 then the group improved and had less errors. If the lower bound was below 1 and the upper bound above 1 then it is unknown if the group had a performance increase or decrease. If the upper and lower bound are more than 1 then the group had an accuracy performance decline and made more damage errors. This data was made using the fixed effect output of the posterior distribution, and can be found on Table G6.

Model Criticism and Model Fit

A detailed analysis for model criticism and fit can be found in Appendix F. The results showed that the gender level held no predictive power. Therefore, gender did not influence the outcome parameters and how performance was estimated by the model. Consequently, gender would not be useful for decision making and adding it as a level to the multilevel model provided no extra information to the posterior distribution. On the other hand, the initial performance level showed predictive power. Therefore, it adds additional information, as we are certain that the groups had notable differences in performance that were overall distinct compared to if initial performance had not been added as a level. In the future this may help create forecasting models that can predict performance at the different sessions (1, 2, 3, 4, 5).

For model fit, it was found that the duration data did not show great fit with any of the distributions tested. The Gamma distribution had the worst fit compared to the ExGaussian and Gaussian distributions, which were comparable to each other. Therefore, when using the ExGaussian distribution, the posterior distribution and the predictions the model made were not as precise as would be expected according to Schmettow (2018). On the other hand, the model fit for the damage count data found that the Poisson distribution had the best fit, the Negative Binomial distribution had the second-best fit, and Binomial distribution had the worst fit. This finding was expected as Schmettow (2018) notes that adding a Poisson distribution to the model should output a posterior distribution where predictions are more precise, as the model more accurately fits the raw data.

Discussion

The main objective of the study is to explore gender and initial performance, and the influence these individual differences have on skill acquisition when using a laparoscopic simulator, known as LapSim. The secondary aim was to use a Bayesian approach for multilevel modelling of learning curves, as it can allow for better fit when modelling complex nonlinear data.

Main Findings

From the research questions there are four key findings. Firstly, for research Question 1 which investigated if gender and initial performance, produced performance outcomes which were transient or enduring. For *gender*, the results from the multilevel model depended upon the performance measure. *Duration* was neither transient nor enduring as both gender groups improved similarly across the sessions, and had similar performance for all sessions (1,2,3,4 and 5). Conversely, a transient effect would have a performance difference between the groups in the first session that disappears with practice. For *damage count* there was a difference between the gender groups in the beginning. Yet, this difference disappeared over time. This lends credibility to the assumption that for gender, accuracy performance is transient. However, inferences based on estimations from the model could not be made to determine what the end outcome was at the last session. Consequently, it cannot be confirmed if gender produced a transient effect for accuracy. On the other hand, results found that initial performance produced transient effects for both duration and damage count. As with practice the groups obtained more similar performance measures. Hence, the findings of the study support the theory by Ackerman (1992), who believed that certain individual differences play less of a substantial role during later stages of skill acquisition, as a task that is practised becomes more procedural and automatic. Furthermore, it is not surprising that no enduring effect was found as visual analysis (see Appendix D, Figure D1 and Figure D2) showed that participants overall individual differences produce performance outcomes that are transient for both duration and damage count. There was more individual variation of scores at the beginning compared to later sessions. Consequently, the results obtained from the multilevel models indicate that certain individual differences – such as initial performance – should not be used for selection processes, as they are not a good criteria for differentiating a good group of future performers from a group of poor performers. The implication of this is that initial one-time testing is not sufficient to get an overall idea of a person's ability and estimating how they will perform in the future.

Secondly, a recommendation by the Institute of Medicine (IOM) (2001, 2012) states there is a necessity to undertake sex-specific reporting. The current study followed this recommendation when investigating Research Question 2 regarding the extent to which gender influenced the learning curve for duration and accuracy performance, as skill acquisition occurred on a laparoscopic simulator. It was found that there was no male group advantage. This is contrary to the literature which leans towards males having an advantage for tasks, such as laparoscopy, which are highly dependent on visuospatial skills (Castro-Alonso & Jansen, 2019; Donnon et al., 2005; Thorson et al., 2011). Although, the Groenier et al. (2015) study had many non-significant gender differences, the one gender difference they did find showed a male advantage with males being more efficient with their movements. Although the current study did not explore this possibility, it was found that there was a female starting advantage in terms of accuracy as the female participants produced a smaller number of errors. However, they did not maintain this advantage with practice. Studies in the medical field have shown that sex differences may not always favour males. For example, female physicians and surgeons show more attentiveness and patients undergoing surgery with a female surgeon have lower mortality rates (Wallis et al., 2017). Though, the current finding that gender differences are not pronounced and dissipate with ongoing practice and training is in line with the skill acquisition model by Dreyfus and Dreyfus (1980) where individuals both reach milestones and improve with practice.

Additionally, for research Question 2, although the female group had a starting advantage for damage count, with practice this advantage subsided. Nevertheless, what is noteworthy is that although they made more mistakes, they became faster at performing the task. Although entirely speculatory, as the current study did not explore a speed-accuracy trade-off, previous studies that do examine the trade-off show similar patterns where bad performance in one measure (e.g. accuracy) creates the opportunity for better execution of another performance measure (e.g. speed). For example, a study by Batmaz, de Mathelin, and Dresp-Langley (2016) showed that people

who worked to increase speed showed a decline in precision. Future research is needed to ascertain if there were trade-offs, and if so, what participants may have chosen to improve or sabotage.

Lastly, for research Question 3 which investigated the extent to which initial performance of duration or damage count, influenced the learning curve of the respective performance measure, as skill acquisition occurred on a laparoscopic simulator. It was found that differences in initial performance became less pronounced with practice. For both duration and damage count, the low initial performing groups improved to a greater extent compared to the high initial performing groups. For duration, this improvement was substantial enough that by session 5, the low initial performing group, for this performance measure, no longer held a disadvantage. This finding goes against the idea that improvement is impossible for a selected few (Alvand et al., 2011; Grantcharov & Funch-Jensen, 2009; Louridas et al., 2017; MacMillan & Cuschieri, 1999; Schijven & Jakimowicz, 2004). On the other hand, the current finding does support other research in the laparoscopic field. For example, a study by Bansal et al. (2012) compared surgeons with exposure to laparoscopy vs. naïve surgeons with no exposure. They found that, although having experience improved duration times in the beginning, by the end of the training period, naïve surgeons and experienced surgeons showed no differences. Furthermore, a key finding is that all participants in the study showed improvement with practice. It is possible that this key finding which investigated each participant's initial baseline score compared to their last scores. As well as the current study that incorporates the individual level and its variation through multilevel modelling. Both do not support the previous findings that performance does not improve for a select few, as it is a myth that although may be found at the group level, is not represented at an individual level. Therefore, it is important when creating prediction models that the individual level is incorporated to obtain a more accurate representation of performance. In conclusion, what the current study and the study by Bansal et al. (2012) indicate is that even with an initial advantage, practice is sufficient in allowing individuals to overcome their starting disadvantages. It is therefore possible to assume that beginning performance is not a good indication of future performance.

Bayesian Approach

A Bayesian approach has benefits compared to a Frequentist approach. Bayesian multilevel modelling, as according to Bürkner (2017) can more appropriately fit learning curve data, which is longitudinal, and typically non-linear.

Many studies have been conducted trying to compare Frequentist multilevel modelling to Bayesian Multilevel modelling. Stegmueller (2013) found that a Bayesian approach had both unbiased estimates and had greater inferential accuracy, compared to Frequentist multilevel modelling that used maximum likelihood (ML) estimates and confidence intervals. However, a later replication study by Elff, Heisig, Schaeffer, and Shikano (2020) noted that a Frequentist approach may not be as biased as Stegmueller (2013) believed. Elff et al. (2020) found that using a restricted maximum likelihood (REML) instead of ML produced unbiased results.

Therefore, as the Groenier et al. (2015) and Groenier et al. (2014) studies both used REML instead of ML, the result they found should be unbiased. Nevertheless, it can be argued the current study still has a more appropriate method. As supported by Browne and Draper (2006) who took both ML and REML into account, but still concluded that the Bayesian approach had more precise estimations.

Multilevel Modelling vs. Theoretical Modelling

Pusic et al. (2017) described various methods available for modelling learning curves and which methods hold greater potential. According to Pusic et al. (2017) linear modelling which although is a standard statistical approach and easy to implement, does a poor job at representing the raw learning curve data. This is a problem as the less representative the model is of the actual data the less likely accurate inferences can be made. On the other hand, Pusic et al. (2017) noted that other techniques, that are more effortful, such as multilevel modelling and theoretical modelling are more suitable, and both hold different benefits depending on what the researcher hopes to achieve. For example, a multilevel model is a statistical method that uses the observed data

directly. It is beneficial as it can refine predictions by using levels, and these predictions can then be used to provide information that can be incorporated into individualized training programs (Pusic et al., 2017). On the other hand, theoretical modelling does not implement levelling, but is a great option as it provides learning parameters and accurate modelling, as it does not use the observed data directly, but instead links the data to a distribution which is assumed to fit how learning occurs as a concept based on theory (Pusic et al., 2017). Statistical analysis is still possible on these models; one such method is to do statistical analysis on the learning parameters (rate, asymptote, amplitude, etc..) that can be outputted from the model (Schröder, Schmettow, & Groenier, 2019).

For theoretical modelling, Heathcote, Brown, & Mewhort (2000) determined that an exponential distribution works best for learning curves. This is because an exponential distribution assumes practice has occurred which improves the likelihood of obtaining a plausible asymptote. Secondly, they give a better fit for unaveraged data sets; for example, individual learning curves whereby each participant gets a separate learning curve. Lastly, it considers a constant learning rate whereby learning slows down based on the stage of learning. The debate for which distribution should be used for modelling theoretical learning is ongoing. Although researchers such as Heathcote et al. (2000) settled upon exponential learning curves, Pusic et al. (2015) believe in modelling based on the power law best represents the learning process.

As currently it is a matter of personal opinion which approach is chosen, it is important that researchers deliberately examine their choices. This is of importance as picking a specific distribution will mean that there will be differing assumptions that may affect the interpretation and inferences that can be obtained from the results. For example, an assumption in an exponential curve is that there is a constant rate of learning which is based on what is left to be learnt. Conversely, a power function assumes there is a mechanism that instead slows down the learning rate (Heathcote et al., 2000).

Decision Making: Bayesian Forecasting and Prediction Modelling using Informative Priors

Prediction modelling is useful as a decision-making tool for objective assessment. It can allow for trainee selection to be based off information learnt from previously gathered data (McElreath, 2018). Therefore, predictions for future participants can be made, both at a group level and even at an individual level (Ni, Groenwold, Nielen, & Klugkist, 2018).

The Bayesian approach compared to the Frequentist approach is advantageous for prediction modelling for two reasons. Firstly, a limitation of the Frequentist approach is that it uses fixed points for parameter estimating, while Bayesian uses prior and posterior distributions (McElreath, 2018). The implementation of distributions is useful in determining the uncertainty of a prediction through a credibility interval. This is a crucial advantage as inferences regarding differences between groups can be viewed as differences in proportions (Phillips, 1975). Therefore, a small difference in proportion indicates the two groups are not different, while a large difference indicates two groups are different (Phillips, 1975). This is useful concept and more informative for decision making as not only can it be determined if a difference is apparent but also "how much of a difference" can be inferred between the two groups (Phillips, 1975). While Frequentist statistics can determine if there is a difference if fails to answer what is the extent of the difference (Phillips, 1975).

Secondly, prediction modelling is more intuitive and more easily integrated using the Bayesian approach (Ni et al., 2018). A key component of the Bayesian theory is that "opinions are expressed as probabilities", a probability known as a prior distribution is chosen, data is then inputted and this data will change the prior distribution to output a posterior distribution (Phillips, 1975). In a Bayesian approach a researcher can decide on an informative or non-informative prior (Zhang et al., 2007). A non-informative prior has model parameters that are not specifically inputted and therefore the prior distribution chosen does not express a specific opinion (Zhang et al., 2007). An informative prior uses past information to formulate an opinion, which is then inputted as

specific parameter values into a model (Zhang et al., 2007). The addition of Informative priors and the expression of an opinion is not easily transferable to a Frequentist approach (Ni et al., 2018). However, it is this ability to easily add informative priors that can be used to create prediction models as shown in research by Zhang et al. (2007) and Ni, Groenwold, Nielen, and Klugkist (2018).

Research by Zhang et al. (2007) chose a method to create predictive models that firstly made a model using non-informative priors, then added observed data to produce a posterior distribution. Information from the posterior distribution was then used as informative priors to create a new prediction model. The study by Ni et al. (2018) goes further and creates informative priors using expert opinion to determine two aspects, how much of the prior distribution should be used for prediction (eg, half, a third, a fifth), and which part of the distribution should be used. Therefore, a weakly informative prior may split the distribution in half, and a selection will take place for which half of the distribution should be used. While for a highly informative prior, only a fifth of the distribution will be selected and used for predictive modelling.

The results found by Zhang et al. (2007) showed that adding an informative prior increased the statistical power of the prediction model, allowing for more accurate predictions. This is a similar conclusion to the Ni et al. (2018) simulation study that found when using a highly informative prior even if the prior was discrepant or incorrect, it still held greater predictive power than if no informative prior was added; in the study a discrepant choice was made when the part of the distribution selected did not have the true value, but rather an adjacent area was purposefully selected. From this finding we assume for our current study that although model criticism found that the duration data did not have a completely accurate model fit in regards to its posterior distribution output, it may still be possible that even if it did not give precise predictions it still may provide informative knowledge to create an enhanced prediction model. The Ni et al. (2018) study also found that although a Frequentist model and a Bayesian model with no informative prior will give the same predictive ability for new data, when even a weak informative prior is added to a Bayesian

model, the predictive ability increases greatly. Furthermore, the more informative the prior the greater the predictive ability will be. From these findings we can conclude that for the current study as the gender groups held no predictive power, it would not be useful as an informative prior, while on the other hand the initial performance groups could provide information that could create a more accurate prediction model.

There are however limitations, and although "fitting a model is easy, prediction is hard" McElreath, 2018). This is because of three reasons, firstly, future data will never fully resemble past data. Secondly, "complex models often make worse predictions than simpler models" (McElreath, 2018). Lastly, making a good model that can be used for prediction is highly dependent on the sample size of the data used to create the prediction model. A small sample size may in fact make the predictive model inaccurate. As the current study only had 67 participants, compared to the Ni et al. (2018) study that had a simulated sample size of 500, and the Zhang et al. (2007) study that had 173 participants. It is consequently possible that the data from the current study may not provide enough information to be used as a prior; and estimating of the parameters and making them more precise, with a small sample size, may lead to more errors than simply ignoring it and providing no prior information at all (McElreath, 2018). However, as already noted a false parameter estimate does not necessarily lead to poor predictions, as Ni et al. (2018) paper shows an incorrect parameter still led to better predictive ability. Nevertheless, this makes prediction even more complicated as incorrect assumptions may either be damaging, have no ill effect at all, or still be beneficial. This is problematic when prediction models are used to obtain answers for high-stake decisions. For example, for minimally invasive surgery a high-stake decision would be deciding if someone would become a surgeon or not. An incorrect decision can be damaging as either someone is denied a career based off incorrect information, or someone is allowed a career but will be incompetent. To make these decisions it is necessary that highly reliable and valid prediction models are used, and although Bayesian and/or multilevel modelling might be more effortful, their

ability to provide more accurate prediction models makes the effort worthwhile for these high-stake decisions (McElreath, 2018; Ni et al., 2018; Pusic et al., 2017).

Limitations

There are four main limitations in the current study. The first limitation is that it does not actually investigate trade-offs and which performance measures were related to one another, and if a performance decline in one measure accounted for the performance improvement in another. Therefore, although duration and damage count as an accuracy measure were analysed, a speed-accuracy trade-off could not be determined with the current method. The second potential limitation is that the bimodal nature of the duration data meant that an ExGaussian distribution which should have according to Schmettow (2018) given a better model fit for time-related data, did not have any benefit over using a Gaussian distribution. Therefore, it is likely the bimodal distribution of the duration data led to inappropriate model fit which could have caused imprecise predictions. The third limitation is that the models that had gender as a level did not show predictive ability (as shown in Appendix F, Table F1 and Table F2). This is important for decision making as it appears gender should not be used for predictive modelling. Therefore, the posterior distribution of the gender models does not change expert opinion, and as an informative prior it would not help predict future trainee performance. Lastly, the initial performance grouping only takes account of one performance measure, rather than a composite score of all performance measures. An overall score in initial performance would have been useful for observing how different measures co-occur with practice, which could not be done with the method used in the current study.

There are an additional three limitations regarding the use of the Bayesian Approach. The first limitation is that the choice of using multilevel models as a statistical approach was limiting as it could not obtain output parameters (e.g. rate, asymptote, amplitude, etc..) which a theoretical model can achieve. Therefore, the current study did not make full use of the Bayesian Approach. This limited the analysis, as firstly, it is not possible to explore how learning parameters may have

been influenced by individual differences and secondly how these different parameters may have influenced each other. The second limitation is that the Bayesian approach is difficult to utilize, inaccessible, and overall it is difficult to put-forth the best approach to publishing results for outside readers who may only have Frequentist knowledge (Anderson, 1998; Gelman, 2008). The third limitation is that processing Bayesian models is time consuming (Browne & Draper, 2006; Ni et al., 2018). One reason is that Bayesian models often use Markov Chain Monte Carlo (MCMC) techniques as they are highly accurate, however as data sets have become larger and more complex, processing has gotten slower (Robert, Elvira, Tawn, & Wu, 2018).

Recommendations

Two types of recommendations are made. The first is for practice and advice for practitioners in the field. The second is suggestions for future exploration.

Practice

Three recommendations for practice can be given regarding individualized training programs, selection and assessment, and lastly training program type.

Firstly, there is no need for specific training programs for gender as there were no substantial gender differences. This contradicts past research from Donnon et al. (2005) who determined that female residency students learning from one-on-one training with feedback created an advantage for acquiring skills and overall preferred method of choice for female students. On the other hand, the finding from the current study supports past research by Saalwachter, Freischlag, Sawyer, and Sanfey (2005) where they observed that surgeons and trainees of both genders have the same objectives regarding what they want in a training program. It is therefore apparent that there is no need for individualized training programs based on gender.

Secondly, individual differences, such as initial performance should not be used for selection or assessment of potential surgical candidates. For example, a novice trainee with low initial laparoscopic surgical performance, should not be judged based on this variable as it is not an

indicator that their performance in the future will still be disadvantaged by having a starting disadvantage. It is therefore recommended that one-time initial testing should be avoided.

Lastly, a proficiency-based program may work for low initial performers as the current study showed that a low performing group has the potential, with practice, to reach the same proficiency level as those with high initial performance. However, they may need more practice or training to achieve this goal as seen in the current study where for duration they only managed to overcome their disadvantage in session 5. For earlier sessions, there was still a disadvantage compared to the high initial performing group.

Future Exploration

Three recommendations for future exploration are given. This included investigating trade-offs, the ranking of high and low performance groups, and implementing Bayesian prediction models.

Firstly, there needs to be research into the individual differences and their influence on specific trade-offs. Typically, a decision-making process takes place when confronted with a task. This can lead to trade-offs when one aspect or performance may be prioritized leading to a decline in another. The speed-accuracy trade-off is acknowledged by Wickelgren (1977) as being of importance to measure. He notes that reaction time studies that focus on speed alone cannot be used to make inferences for a task, such as if the task was easy or if a participant performed well overall. For example, at session 2, for the grasping task, participant 52 and 23 had similar median finishing times (37.72 and 38.71 seconds, respectively). Based off speed alone it may be assumed that participants are equal, however, participant 52 had a median damage count of 8, while participant 23 only had a median damage count of 4. The addition of an accuracy score greatly changes inferences for how the participants performed in the task; and therefore, this changes our ultimate understanding of how difficult the task was for each participant; as well as which participants (in this case participant 52) are having difficulties and may benefit from the implementation of an individualized training program.

Secondly, instead of using quartiles for a specific measure, a composite score that accounted for all performance measures could have been used. This would have been a better way to rank high or low performers and determine those who may have needed individualized training programs; as well as a more practical method to use for exploring future research into trade-offs. However, it is difficult to obtain a composite measure as there are many different variables all with different scales and units, nonetheless the use of weights and/or Bayesian hierarchical latent variable modelling (BLVM) have shown to be promising methods to obtain composite scores (Iyengar et al., 2019; Shwartz et al., 2008; Staiger, Dimick, Baser, Fan, & Birkmeyer, 2009).

Thirdly, the Bayesian approach is unique in that it can integrate the creation of prediction models, using informative priors that are established from past data and learning. Such modelling allows the ability to forecast and predict how a future participant might perform as they acquire a skill on LapSim, making it useful for assessing potential candidates. Additionally, the ability to make better decisions could further be implemented in the future to predict if an individual would benefit from an individualized training program. Research by Hooten, Johnson, and Brost (2019) introduced a new method for multilevel modelling that uses Prior- Recursive Bayes and Proposal-Recursive Bayes to fit data as it becomes available. In the future, this technique may have an advantage as it can implement forecasting and predictions continuously, as well allow faster updating of a model compared to current Markov Chain Monte Carlo (MCMC) techniques. However, two suggestions are relevant to create accurate predictions. Firstly, a large sample size is needed, and secondly using models that are overall less complex may be necessary for accurate predictions (McElreath, 2018).

Conclusion

By modelling skill acquisition for a laparoscopic task using a virtual reality simulator, it was found that although individual differences may cause more variation in performance in the beginning, with training and more practice, these differences do not have as much influence. This also alludes to a distinct possibility that even low initial performers with practice and time have the potential to overcome the steep learning curve in laparoscopy surgery. Modelling and estimating

what a person will do in the future based on a limited context is not possible. Therefore, one-time testing does not give a great representation of potential future performance. This study could be seen as a clear indication of why not to "judge a book by its cover" - as initial presumptions such as a male participant will have a surgical advantage and a low initial performer will not improve are statements that do not show merit in the laparoscopic field. However, more research is needed to understand decision making processes and possible trade-offs. Once this is established, it should be possible to determine if specialized training programs are necessary or if resources could better be used elsewhere. Nevertheless, already assumptions can be formulated, for example, the current study indicated that specialized training may not be needed for gender differences, and resources may be better allocated to other programs.

The implications of the current study are as follow. Firstly, it is recommended that one-time testing and screening is inappropriate and should be avoided when selecting and identifying potential trainees as individual differences can be overcome with practice. Secondly, individualized training programs for gender are not recommended, as differences between the groups are not pronounced enough to warrant their need. This goes against past research by Donnon et al. (2005) who suggested laparoscopic skill acquisition was different for males and females, and that the implementation of one-on-one training was advantageous for female trainees. Lastly, there is no male advantage when using the laparoscopic simulator. This is an exciting discovery, as the tasks in the study had a high visuospatial dependency. Consequently, the finding does not confirm past research which has indicated that male participants have an advantage for these types of visuospatial tasks (Donnon et al., 2005; Grantcharov et al., 2003; Thorson et al., 2011).

References

Ackerman, P. L. (1992). Predicting Individual Differences in Complex Skill Acquisition: Dynamics of Ability Determinants. *Journal of Applied Psychology*, *77*(5), 598–614. https://doi.org/10.1037/0021-9010.77.5.598

Adams, J. A. (1957). The relationship between certain measures of ability and the acquisition of a psychomotor criterion response. *The Journal of General Psychology*. https://doi.org/10.1080/00221309.1957.9918366

Ahlberg, G., Enochsson, L., Gallagher, A. G., Hedman, L., Hogman, C., McClusky, D. A., … Arvidsson, D. (2007). Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. *American Journal of Surgery*. https://doi.org/10.1016/j.amjsurg.2006.06.050

Alvand, A., Auplish, S., Khan, T., Gill, H. S., & Rees, J. L. (2011). Identifying orthopaedic surgeons of the future - The inability of some medical students to achieve competence in basic arthroscopic tasks despite training: A randomised study. *Journal of Bone and Joint Surgery - Series B*. https://doi.org/10.1302/0301-620X.93B12.27946

Anderson, J. (1998). Embracing uncertainty: The interface of Bayesian statistics and cognitive psychology. *Ecology and Society*, *2*(1). https://doi.org/10.5751/es-00043-020102

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*. https://doi.org/10.1037/0033-295X.89.4.369

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of Foreign Language Vocabulary and the Spacing Effect. *Psychological Science*. https://doi.org/10.1111/j.1467-9280.1993.tb00571.x

Bansal, V. K., Tamang, T., Misra, M. C., Prakash, P., Rajan, K., Bhattacharjee, H. K., … Goswami, A. (2012). Laparoscopic suturing skills acquisition: A comparison between laparoscopy-exposed

and laparoscopy-naive surgeons. *Journal of the Society of Laparoendoscopic Surgeons*, *16*(4), 623–631. https://doi.org/10.4293/108680812X13462882737375

Batmaz, A. U., de Mathelin, M., & Dresp-Langley, B. (2016). Getting nowhere fast: Trade-off between speed and precision in training to execute image-guided hand-tool movements. *BMC Psychology*, *4*(1), 1–19. https://doi.org/10.1186/s40359-016-0161-0

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*. https://doi.org/10.1214/088342304000000116

Becker, J., Berkley, K., Geary, N., Hampson, E., Herman, J., & Young, E. (Eds.). (2007). *Sex Differences in the Brain: From Genes to Behavior*. Retrieved from https://books.google.nl/books?id=IeaLXPWsbuAC&pg=PR14&lpg=PR14&dq=institute+of+medicine+IOM+gender+exploration+learning&source=bl&ots=7JRE3GU-52&sig=ACfU3U3zyHeVUjah4Em-lPFpJbMBIWJkzg&hl=en&sa=X&ved=2ahUKEwiI9b6KuszqAhVG2KQKHSshCGwQ6AEwBHoECAwQAQ#v=onepag

Bennett, C. L., Stryker, S. J., Rosario Ferreira, M., Adams, J., & Bert, R. W. (1997). The learning curve for laparoscopic colorectal surgery: Preliminary results from a prospective analysis of 1194 laparoscopic-assisted colectomies. *Archives of Surgery*. https://doi.org/10.1001/archsurg.1997.01430250043009

Bissolati, M., Orsenigo, E., & Staudacher, C. (2016). Minimally invasive approach to colorectal cancer: an evidence-based analysis. *Updates in Surgery*. https://doi.org/10.1007/s13304-016-0350-7

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*. https://doi.org/10.1214/06-BA117

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v080.i01

Castro-Alonso, J. C., & Jansen, P. (2019). Sex Differences in Visuospatial Processing. In *Visuospatial Processing for Education in Health and Natural Sciences* (pp. 81–110). https://doi.org/10.1007/978-3-030-20969-8_4

Donnon, T., DesCôteaux, J. G., & Violato, C. (2005). Impact of cognitive imaging and sex differences on the development of laparoscopic suturing skills. *Canadian Journal of Surgery*, *48*(5), 387–393.

Dreyfus, S. E., & Dreyfus, H. L. (1980). *A Five-Stage Model of Mental Activites Involved in Directed Skill Acquisition*.

Elff, M., Heisig, J. P., Schaeffer, M., & Shikano, S. (2020). Multilevel Analysis with Few Clusters: Improving Likelihood-Based Methods to Provide Unbiased Estimates and Accurate Inference. *British Journal of Political Science*. https://doi.org/10.1017/S0007123419000097

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine*. https://doi.org/10.1097/00001888-200410001-00022

Fitts, P. M. (1964). Perceptual-Motor Skill Learning. *Categories of Human Learning*. https://doi.org/10.1080/00140139208967796

Gabry, J., & Goodrich, B. (n.d.). rstanarm: prior distributions and options. Retrieved June 9, 2020, from https://mc-stan.org/rstanarm/reference/priors.html#references

Galaal, K., Bryant, A., Fisher, A. D., Al-Khaduri, M., Kew, F., & Lopes, A. D. (2012). Laparoscopy versus laparotomy for the management of early stage endometrial cancer. *Cochrane Database of Systematic Reviews*. https://doi.org/10.1002/14651858.CD006655.pub2

Gallagher, A., McClure, N., McGuigan, J., Ritchie, K., & Sheehy, N. (1998). An ergonomic analysis of the fulcrum effect in the acquisition of endoscopic skills. *Endoscopy*. https://doi.org/10.1055/s-

2007-1001366

Gallagher, A., Ritter, M., Champion, H., Higgins, G., Fried, M., Moses, G., … Satava, R. (2005). Virtual reality simulation for the operating room: Proficiency-based training as a paradigm shift in surgical skills training. *Annals of Surgery*. https://doi.org/10.1097/01.sla.0000151982.85062.80

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*. https://doi.org/10.1214/08-BA318

Grantcharov, T. P., Bardram, L., Funch-Jensen, P., & Rosenberg, J. (2003). Impact of hand dominance, gender, and experience with computer games on performance in virtual reality laparoscopy. *Surgical Endoscopy and Other Interventional Techniques*, *17*(7), 1082–1085. https://doi.org/10.1007/s00464-002-9176-0

Grantcharov, T. P., & Funch-Jensen, P. (2009). Can everyone achieve proficiency with the laparoscopic technique? Learning curve patterns in technical skills acquisition. *American Journal of Surgery*. https://doi.org/10.1016/j.amjsurg.2008.01.024

Groenier, M., Groenier, K. H., Miedema, H. A. T., & Broeders, I. A. M. J. (2015). Perceptual Speed and Psychomotor Ability Predict Laparoscopic Skill Acquisition on a Simulator. *Journal of Surgical Education*. https://doi.org/10.1016/j.jsurg.2015.07.006

Groenier, M., Schraagen, J. M. C., Miedema, H. A. T., & Broeders, I. A. J. M. (2014). The role of cognitive abilities in laparoscopic simulator training. *Advances in Health Sciences Education*. https://doi.org/10.1007/s10459-013-9455-7

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/BF03212979

Hoffman, L. (2007). Multilevel Models for Examining Individual Differences in Within-Person

Variation and Covariation Over Time. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273170701710072

Hooten, M. B., Johnson, D. S., & Brost, B. M. (2019). Making Recursive Bayesian Inference Accessible. *American Statistician*. https://doi.org/10.1080/00031305.2019.1665584

Institute of Medicine (IOM). (2001). *Exploring the biological contributions to human health: Does sex matter?*

Institute of Medicine (IOM). (2012). *Sex-specific reporting of scientific research: A workshop summary*. Washington, DC: The National Academies Press.

Iyengar, B. R. S., Mossner, B. J. M., Sekhri, B. S., Mullard, M. A., Krapohl, P. G., Campbell, M. D. A., & Englesbe, M. M. J. (2019). A New Composite Measure for Assessing Surgical Performance. *Michigan Journal of Medicine*. https://doi.org/10.3998/mjm.13761231.0004.113

James, W. (1891). The Principles of Psychology. *The American Journal of Psychology*. https://doi.org/10.2307/1412102

Keehner, M., Lippa, Y., Montello, D. R., Tendick, F., & Hegarty, M. (2006). Learning a spatial skill for surgery: how the contributions of abilities change with practice. *Applied Cognitive Psychology*, *20*(4), 487–503. https://doi.org/10.1002/acp.1198

Kolkman, W., Wolterbeek, R., & Jansen, F. W. (2005). Gynecological laparoscopy in residency training program: Dutch perspectives. *Surgical Endoscopy and Other Interventional Techniques*, *19*(11), 1498–1502. https://doi.org/10.1007/s00464-005-0291-6

Louridas, M., Szasz, P., Fecso, A. B., Zywiel, M. G., Lak, P., Bener, A. B., … Grantcharov, T. P. (2017). Practice does not always make perfect: need for selection curricula in modern surgical training. *Surgical Endoscopy*. https://doi.org/10.1007/s00464-017-5572-3

MacMillan, A. I. M., & Cuschieri, A. (1999). Assessment of innate ability and skills for endoscopic

manipulations by the advanced dundee endoscopic psychomotor tester: Predictive and concurrent validity. *American Journal of Surgery*. https://doi.org/10.1016/S0002-9610(99)00016-1

McElreath, R. (2018). Statistical rethinking: A bayesian course with examples in R and stan. In *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. https://doi.org/10.1201/9781315372495

Moorthy, K., Munz, Y., Sarker, S. K., & Darzi, A. (2003). Objective assessment of technical skills in surgery. *British Medical Journal*. https://doi.org/10.1136/bmj.327.7422.1032

Ni, H., Groenwold, R. H. H., Nielen, M., & Klugkist, I. (2018). Prediction models for clustered data with informative priors for the random effects: A simulation study. *BMC Medical Research Methodology*. https://doi.org/10.1186/s12874-018-0543-5

Perez-Cruet, M. J., Fessler, R. G., & Perin, N. I. (2002). Review: Complications of minimally invasive spinal surgery. *Neurosurgery*. https://doi.org/10.1097/00006123-200211002-00005

Phillips, L. D. (1975). Bayesian Statistics for Social Scientists. *Operational Research Quarterly (1970-1977)*, *26*(1), 113. https://doi.org/10.2307/3007832

Pusic, M. V., Boutis, K., Hatala, R., & Cook, D. A. (2015). Learning Curves in Health Professions Education. *Academic Medicine*. https://doi.org/10.1097/ACM.0000000000000681

Pusic, M. V., Boutis, K., Pecaric, M. R., Savenkov, O., Beckstead, J. W., & Jaber, M. Y. (2017). A primer on the statistical modelling of learning curves in health professions education. *Advances in Health Sciences Education*. https://doi.org/10.1007/s10459-016-9709-2

Rieder, E., Martinec, D. V., Cassera, M. A., Goers, T. A., Dunst, C. M., & Swanstrom, L. L. (2011). A triangulating operating platform enhances bimanual performance and reduces surgical workload in single-incision laparoscopy. *Journal of the American College of Surgeons*.

https://doi.org/10.1016/j.jamcollsurg.2010.10.009

Robert, C. P., Elvira, V., Tawn, N., & Wu, C. (2018). Accelerating MCMC algorithms. *Wiley Interdisciplinary Reviews: Computational Statistics*. https://doi.org/10.1002/wics.1435

Rosser, J. C., Rosser, L. E., & Savalgi, R. S. (1997). Skill acquisition and assessment for laparoscopic surgery. *Archives of Surgery*. https://doi.org/10.1001/archsurg.1997.01430260098021

Saalwachter, A. R., Freischlag, J. A., Sawyer, R. G., & Sanfey, H. A. (2005). The training needs and priorities of male and female surgeons and their trainees. *Journal of the American College of Surgeons*. https://doi.org/10.1016/j.jamcollsurg.2005.03.016

Schijven, M. P., & Jakimowicz, J. (2004). The learning curve on the Xitact LS 500 laparoscopy simulator: Profiles of performance. *Surgical Endoscopy and Other Interventional Techniques*. https://doi.org/10.1007/s00464-003-9040-x

Schmettow, M. (2018). New statistics for the design researcher. Retrieved December 31, 2019, from https://schmettow.github.io/New_Stats/GLM.html#rating-scales

Schröder, T., Schmettow, M., & Groenier, M. (2019). *The Influence of time pressure on MIS simulation tasks* (University of Twente). Retrieved from https://essay.utwente.nl/79941/

Seymour, N. E., Gallagher, A. G., Roman, S. A., O'Brien, M. K., Bansal, V. K., Andersen, D. K., … Blumgart, L. H. (2002). Virtual reality training improves operating room performance results of a randomized, double-blinded study. *Annals of Surgery*. https://doi.org/10.1097/00000658-200210000-00008

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*. https://doi.org/10.1037/0033-295X.84.2.127

Shwartz, M., Ren, J., Peköz, E. A., Wang, X., Cohen, A. B., & Restuccia, J. D. (2008). Estimating a

Composite Measure of Hospital Quality From the Hospital Compare Database. *Medical Care*, *46*(8), 778–785. https://doi.org/10.1097/MLR.0b013e31817893dc

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*. https://doi.org/10.1177/0956797611417632

Smith, A. F. M., & Gelfand, A. E. (1992). Bayesian Statistics without Tears: A Sampling-Resampling Perspective. In *Source: The American Statistician* (Vol. 46).

Staiger, D. O., Dimick, J. B., Baser, O., Fan, Z., & Birkmeyer, J. D. (2009). Empirically derived composite measures of surgical performance. *Medical Care*. https://doi.org/10.1097/MLR.0b013e3181847574

Stan Development Team. (2020). RStan: the R interface to Stan. Retrieved from R package version 2.21.2 website: http://mc-stan.org/

Stefanidis, D., Acker, C. E., Swiderski, D., Heniford, B. T., & Greene, F. L. (2008, January 1). Challenges During the Implementation of a Laparoscopic Skills Curriculum in a Busy General Surgery Residency Program. *Journal of Surgical Education*, Vol. 65, pp. 4–7. https://doi.org/10.1016/j.jsurg.2007.11.009

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and bayesian approaches. *American Journal of Political Science*. https://doi.org/10.1111/ajps.12001

Surgical Science. (2012a). LapSim Basic Skills: Grasping. Retrieved June 11, 2020, from https://www.youtube.com/watch?v=LjX9cDnroi8

Surgical Science. (2012b). LapSim Basic Skills: Instrument Navigation. Retrieved June 11, 2020, from https://www.youtube.com/watch?v=IgI5MZFbc4o

Thorson, C. M., Kelly, J. P., Forse, R. A., & Turaga, K. K. (2011). Can we continue to ignore gender differences in performance on simulation trainers? *Journal of Laparoendoscopic and Advanced Surgical Techniques*, *21*(4), 329–333. https://doi.org/10.1089/lap.2010.0368

Van Dongen, K. W., Tournoij, E., Van Der Zee, D. C., Schijven, M. P., & Broeders, I. A. M. J. (2007). Construct validity of the LapSim: Can the LapSim virtual reality simulator distinguish between novices and experts? *Surgical Endoscopy and Other Interventional Techniques*. https://doi.org/10.1007/s00464-006-9188-2

Wallis, C. J., Ravi, B., Coburn, N., Nam, R. K., Detsky, A. S., & Satkunasivam, R. (2017). Comparison of postoperative outcomes among patients treated by male and female surgeons: A population based matched cohort study. *BMJ (Online)*. https://doi.org/10.1136/bmj.j4366

Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. *Acta Psychologica*. https://doi.org/10.1016/0001-6918(77)90012-9

Zhang, Z., Hamagami, F., Lijuan Wang, L., Nesselroade, J. R., & Grimm, K. J. (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development*, *31*(4), 374–383. https://doi.org/10.1177/0165025407077764

Zyphur, M. J., Kaplan, S. A., Islam, G., Barsky, A. P., & Franklin, M. S. (2008). Conducting multilevel analyses in medical education. *Advances in Health Sciences Education*. https://doi.org/10.1007/s10459-007-9078-y

Appendix A

Table A1

Overview of R Libraries and Functions Used in the Current Study

Objective	R Libraries	Functions (input and output)	Notes
Quartile Division		quantile	This was used to create four mostly equally sized groups. One group was high initial performers and the other low initial performers.
Graphical Representation of Learning Curves	tidyverse	ggplot	
Histograms	tidyverse	ggplot	
Violin Plots for Variation	tidyverse	ggplot	
Dispersion Test	AER	glm dispersiontest	A Poisson distribution was used.
Determine individual noise	rstanarm	stan_glmer coef(*model*) ranef(*model*)	Coef was used to obtain overall individual noise. Ranef was used to obtain noise for each participant.
Multilevel Models for Duration	brms	brm fixef(*model*)	An ExGaussian distribution was used. Fixef gives the fixed effects of a model which is obtained from the posterior distribution.
Multilevel Models for Damage Count	rstanarm	stan_glmer fixef(*model*, mean.func = exp)	Fixef gives the fixed effects of a model which is obtained from the posterior distribution. An exponential mean function needs to be used with fixef.
Predictive Power	brms rstanarm	brm Stan_glmer Fixef(*model*) Grpef(*model*)	Fixef gives the fixed effect of a model. Grpef provides sigma which is the variation of posterior distribution.
Model Fit for Duration Data	tidyverse brms	ggplot brm posterior post_pred	Created models for Gaussian, and Gamma distributions
Model fit for Damage Count Data	MASS vcd	displot	Create Q-Q plots for different types of distributions, e.g. Binomial, negative binomial, and Poisson

Note. Shows the main R libraries used and the objective for using them. This information was provided for both replication purposes as well as for transparency.

Appendix B

This section has the descriptive statistics for the different groups (female group, male group, low initial performance groups, and high initial performance groups) and their performance outcomes for both duration and damage count for each session. Table B1, B2, B3 and B4, include the median, mean, standard deviation, minimum value and maximum value.

Table B1

Duration Performance for the Gender Groups based off all Observations in a Session

Session	Gender									
	Female Group (n = 38)					Male Group (n = 29)				
	Mdn	*M*	*SD*	*Min*	*Max*	*Mdn*	*M*	*SD*	*Min*	*Max*
1	41.90	39.62	14.63	5.91	79.53	40.28	38.38	14.34	8.63	60.25
2	43.21	41.43	10.6	5.62	58.58	40.74	38.12	10.16	7.63	58.40
3	39.03	33.57	13.59	4.47	51.85	34.65	30.61	12.98	5.13	60.25
4	20.15	23.51	12.18	4.52	55.48	18.31	23.01	12.76	3.91	53.90
5	17.00	16.81	4.83	3.57	37.09	16.19	15.89	5.67	3.02	34.99

Note. Table shows the group median (*Mdn*), mean (*M*), standard deviation (*SD*), minimum value (*Min*), and maximum value (*Max*), for each session. The number of participants (n) is also given.

Table B2

Duration Performance for the Initial Performance Groups based off all Observations in a Session

Session	Initial Performance									
	High Initial Performance Group (n = 16)					Low Initial Performance Group (n = 16)				
	Mdn	*M*	*SD*	*Min*	*Max*	*Mdn*	*M*	*SD*	*Min*	*Max*
1	31.21	29.76	10.41	5.91	53.75	51.21	49.68	13.11	13.35	79.53
2	39.06	37.52	8.81	5.62	53.52	46.00	44.28	10.32	8.96	56.92
3	36.07	31.06	12.61	4.47	50.32	41.39	35.81	13.98	6.58	51.85
4	16.00	20.01	12.07	3.91	46.73	25.61	28.79	12.06	8.44	55.48
5	15.13	14.9	4.26	3.57	32.64	17.95	18.21	5.311	8.24	37.09

Note. Table shows the group median (*Mdn*), mean (*M*), standard deviation (*SD*), minimum value (*Min*), and maximum value (*Max*), for each session. The number of participants (n) is also given.

Table B3

Damage Count Performance for the Gender Groups based off all Observations in a Session

Session	Gender									
	Female Group (n = 38)					Male Group (n = 29)				
	Mdn	M	SD	Min	Max	Mdn	M	SD	Min	Max
1	3.00	4.81	4.75	0.00	22.00	4.00	4.91	5.10	0.00	22.00
2	5.00	6.88	6.08	0.00	28.00	4.00	6.62	6.06	0.00	25.00
3	5.00	6.23	5.64	0.00	24.00	3.00	5.95	5.71	0.00	24.00
4	3.00	4.14	4.15	0.00	20.00	3.00	3.77	4.07	0.00	22.00
5	3.00	3.29	2.66	0.00	13.00	2.00	2.60	2.55	0.00	12.00

Note. Table shows the group median (*Mdn*), mean (*M*), standard deviation (*SD*), minimum value (*Min*), and maximum value (*Max*), for each session. The number of participants (n) is also given.

Table B4

Damage Count Performance for the Initial Performance Groups based off all Observations in a Session

Session	Initial Performance									
	High Initial Performance Group (n = 16)					Low Initial Performance Group (n = 16)				
	Mdn	*M*	*SD*	*Min*	*Max*	*Mdn*	*M*	*SD*	*Min*	*Max*
1	1.00	1.61	1.52	0.00	5.00	8.00	9.50	5.86	0.00	22.00
2	4.00	5.93	5.80	0.00	24.00	8.00	8.70	6.93	0.00	28.00
3	4.00	5.66	5.17	0.00	21.00	4.00	7.07	6.57	0.00	24.00
4	2.00	3.25	3.30	0.00	16.00	4.00	4.75	3.78	0.00	21.00
5	2.00	2.53	2.26	0.00	10.00	4.00	3.81	2.65	0.00	12.00

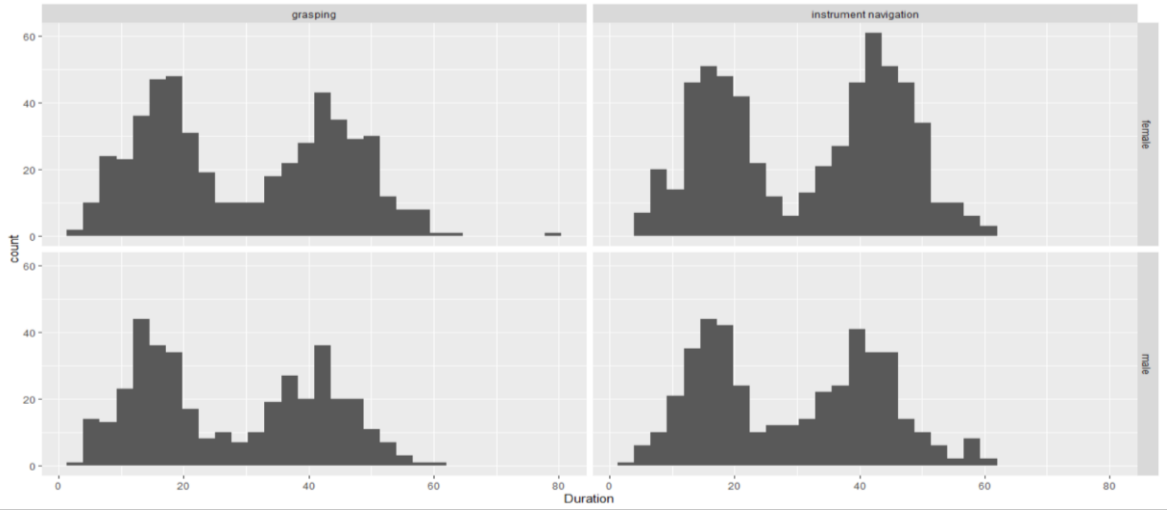
Note. Table shows the group median (*Mdn*), mean (*M*), standard deviation (*SD*), minimum value (*Min*), and maximum value (*Max*), for each session. The number of participants (n) is also given.

Appendix C

The following section holds the figures and analysis relevant for data exploration.

Figure C1

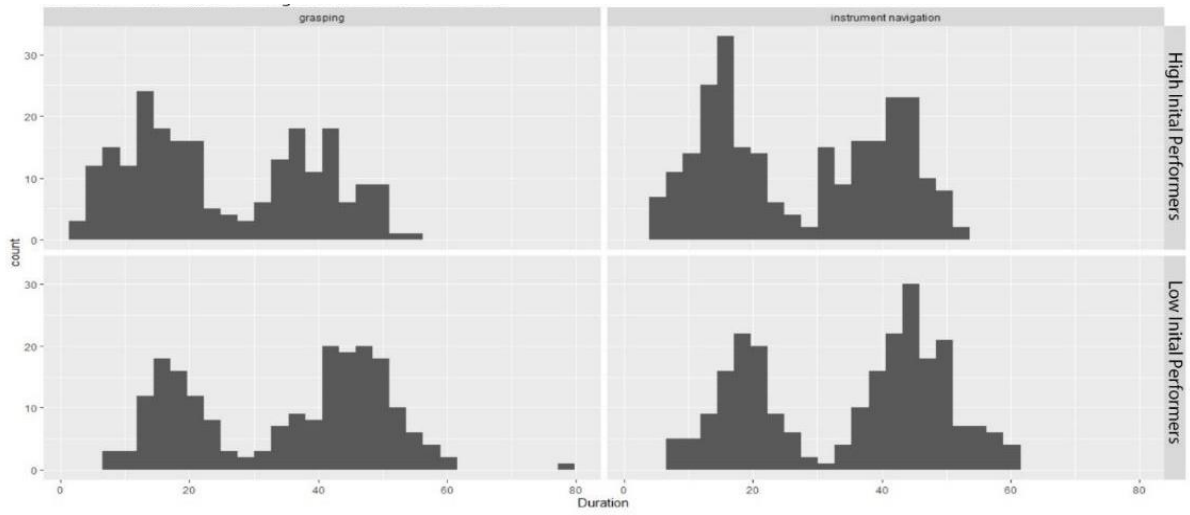
Histogram of Duration for Gender Groups and Task



Note. There is a bimodal peak for both groups and tasks. There are also signs of overdispersion.

Figure C2

Histogram of Duration for Initial Performance Group and Task

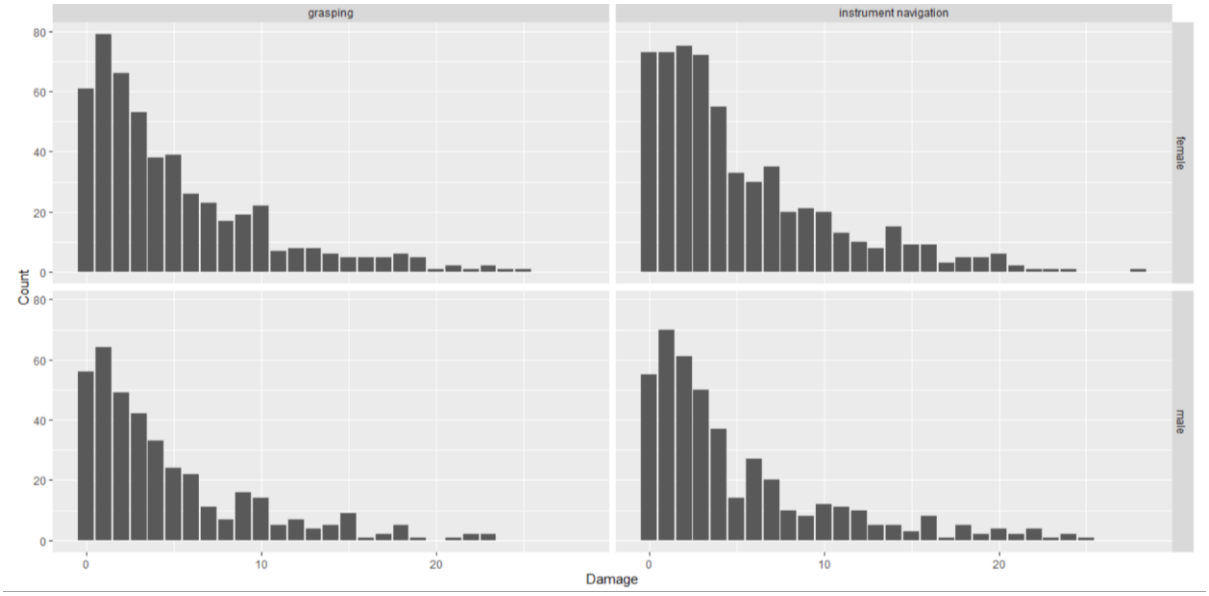


Note. There is a bimodal peak for both groups and tasks. There are also signs of overdispersion.

For damage count data an overdispersion test further verified that overdispersion was apparent, and therefore, the model for this data will have to be modified.

Figure C3

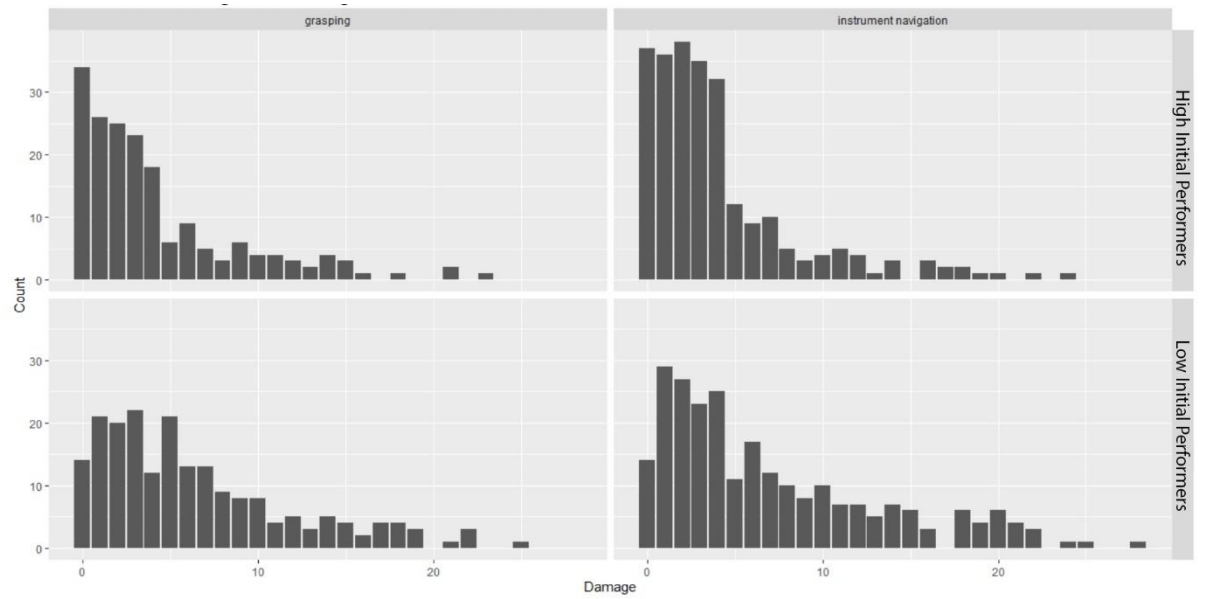
Histogram of Damage Count for Gender Groups and Task



Note. The data is not normally distributed, although there is a unimodal peak. There is a right-skewed distribution with most values being on the left. There are also signs of overdispersion.

Figure C4

Histogram of Damage Count for Initial Performance Groups and Task



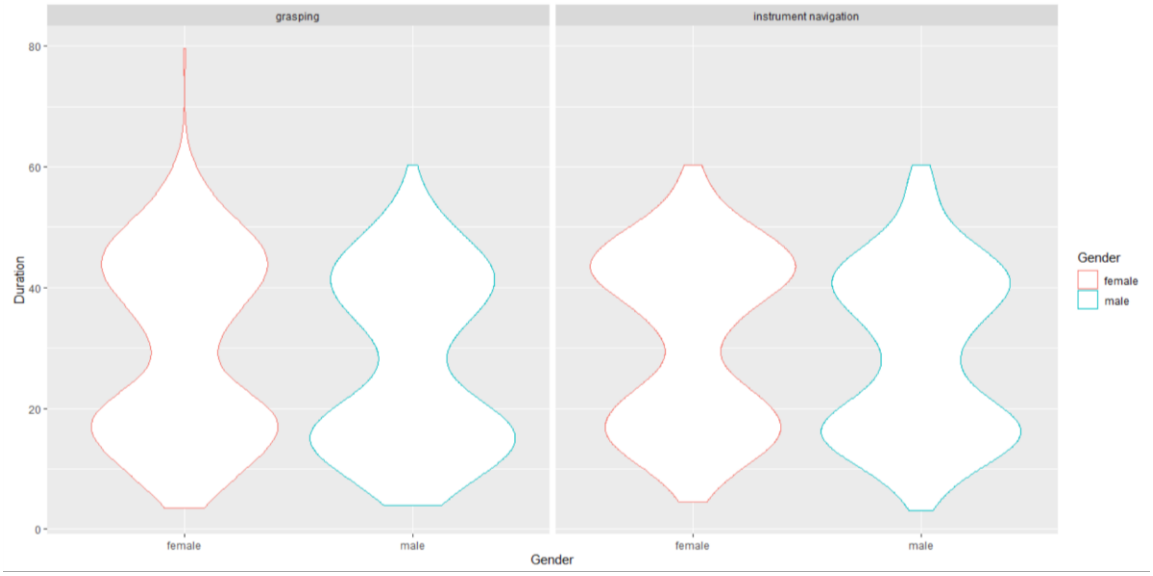
Note. The data is not normally distributed, although there is a unimodal peak. There is a right-skewed distribution with most values being on the left. There are also signs of overdispersion.

Variation was seen using the violin plots. It was observed that females and those with low initial performance displayed more variation. For duration, it was the instrumental navigation task

that showed more variation (Figures C5 & C6). While for damage count, it was the grasping task that showed more variation for these participants (Figures C7 & C8). Although these differences in variation were not considerable enough to affect the type of distributions (e.g. right-tailed, bimodal) as seen in the histograms (Figures C1, C2, C3, & C4). It can therefore be assumed that differences between the groups and tasks, although visible in the violin plots, should not be substantial enough to have a considerable influence on the results.

Figure C5

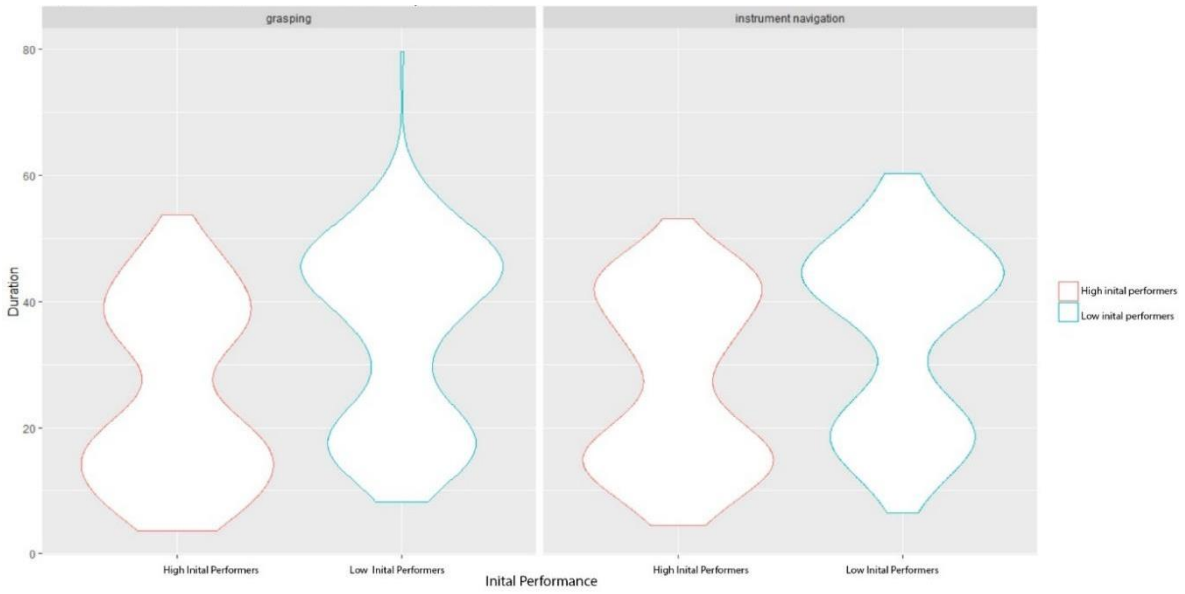
Variation of Duration for Gender Groups and Task



Note. Overall, the female group showed more variation for the grasping task where they had a high max value compared to the male group who showed less variation for this task.

Figure C6

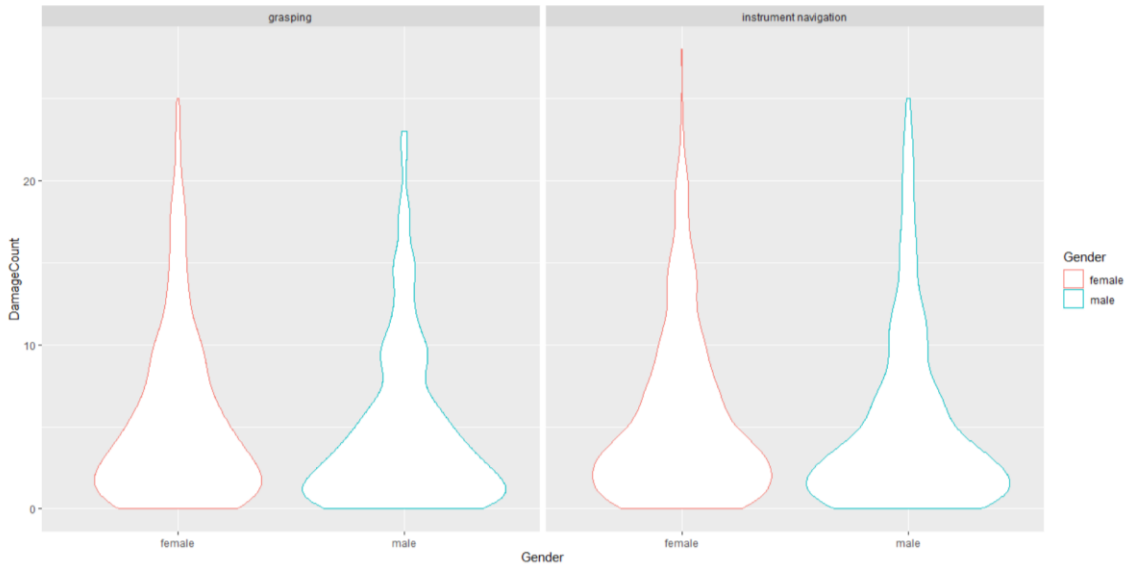
Variation of Duration for Initial Performance Groups and Task



Note. Overall, the low initial performance group showed more variation for both tasks. However, the grasping task showed more variation in the distribution compared to instrument navigation.

Figure C7

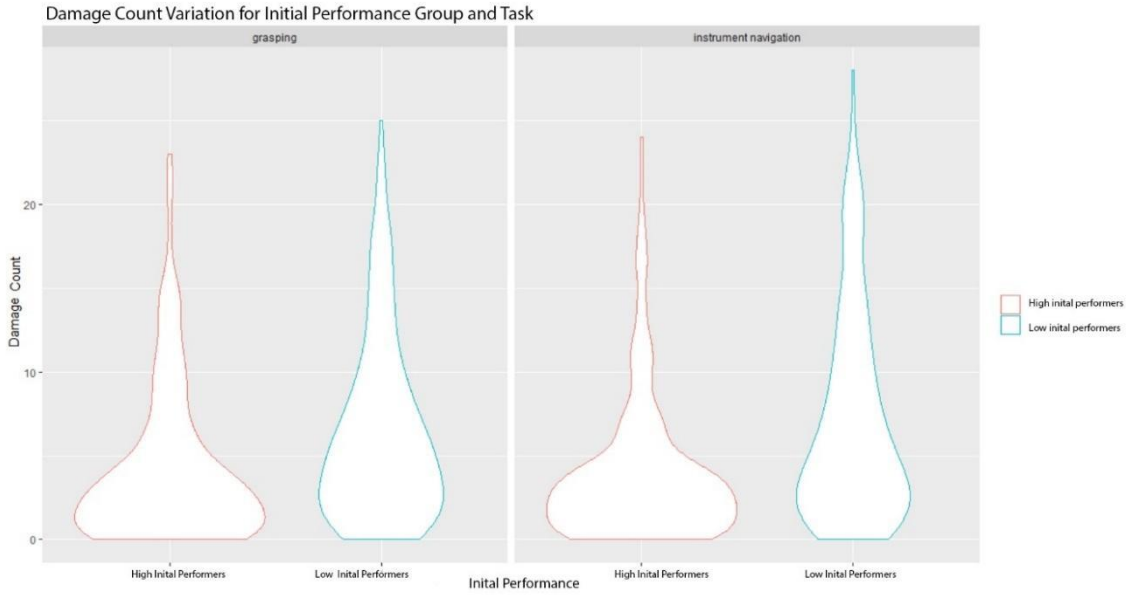
Variation of Damage Count for Gender Groups and Task



Note. The Female group showed more variation in their damage count compared to the male group, for both tasks. However, for instrument navigation, this is slightly more pronounced with female participants showing even more distributed scores.

Figure C8

Variation of Damage Count for Initial Performance Groups and Task



Note. Regardless of the task, it appears that the high initial performing group had less variance compared to the low initial performing group.

Appendix D

The following section holds the analysis relevant for multilevel exploration.

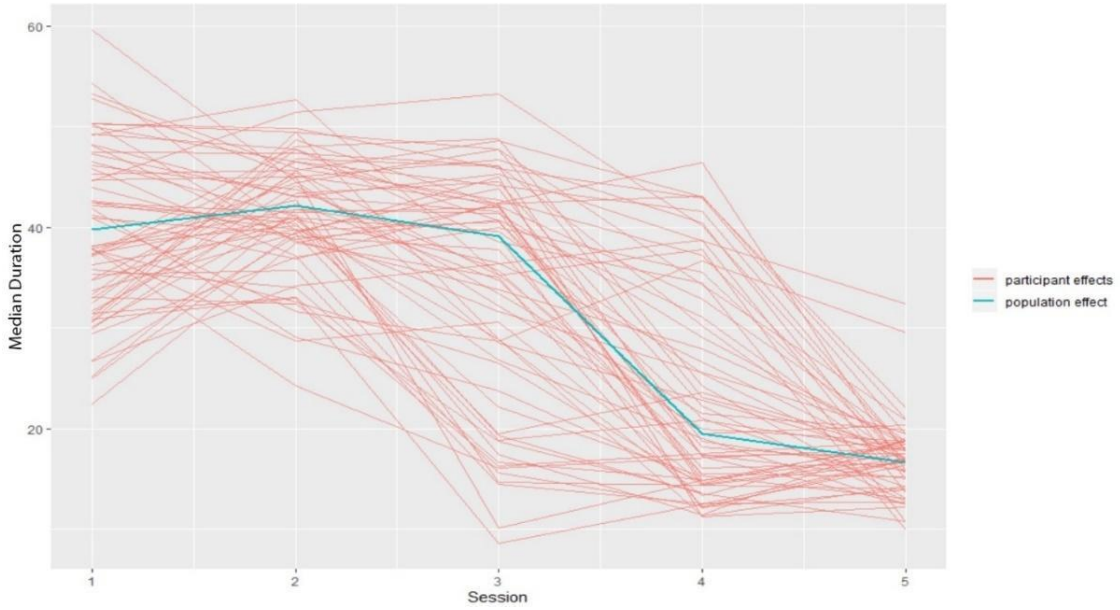
Participant vs. Population Effect

It appears that individuals had more variation (participant effect) for scores at the beginning compared to later sessions. Therefore, overall differences between participants produce transient performance outcomes for both duration and damage count.

For duration, by session 5, there appears to be minimal impact caused by participants individual differences. This can be seen by visual inspection of Figure D1. For sessions 1,2,3 and 4, many individual learning curves do not closely match the population learning curve and, instead, had more spread-out scores. However, by session 5, all individuals have more similar median scores.

Figure D1

Individual Level Duration Learning Curves (Participant Effect) with Group Level Learning Curve (Population Effect)

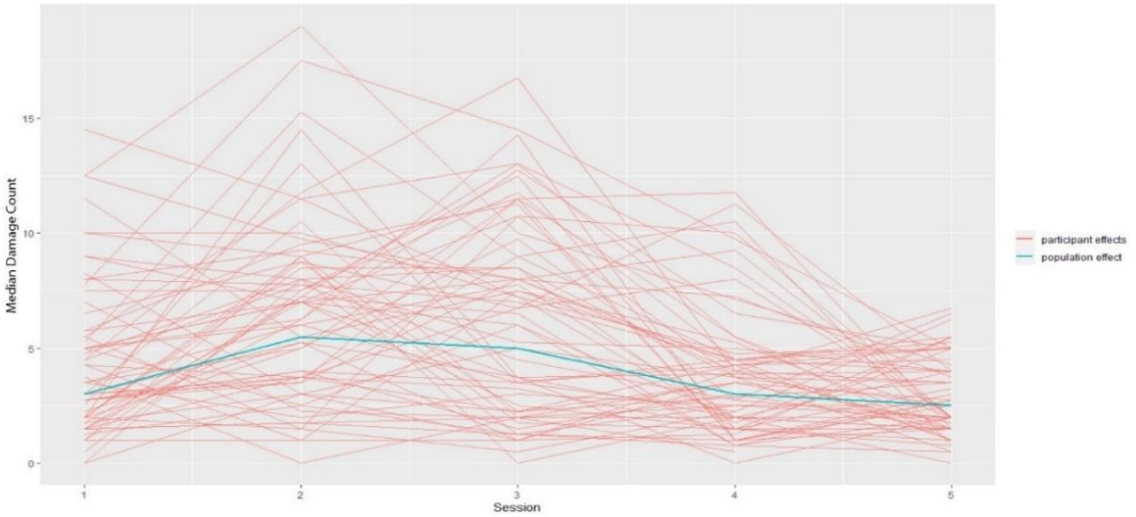


Note. The blue line is the median for all participants (population effect), while the red lines show how each participant varied (participant effect) and what median measure they obtained for each session.

As with duration, a similar result is seen for damage count. In Figure D2, at session 1 and 5, damage count performance was the same at the population level. However, on the participant level, the scores become denser and converge at session 5.

Figure D2

Individual Level Damage Count Learning Curves (Participant Effect) with Group Level Learning Curve (Population Effect)



Note. The blue line is the median for all participants (population effect), while the red lines show how each participant varied (participant effect) and what median measure they obtained for each session.

Individual Learning Curves

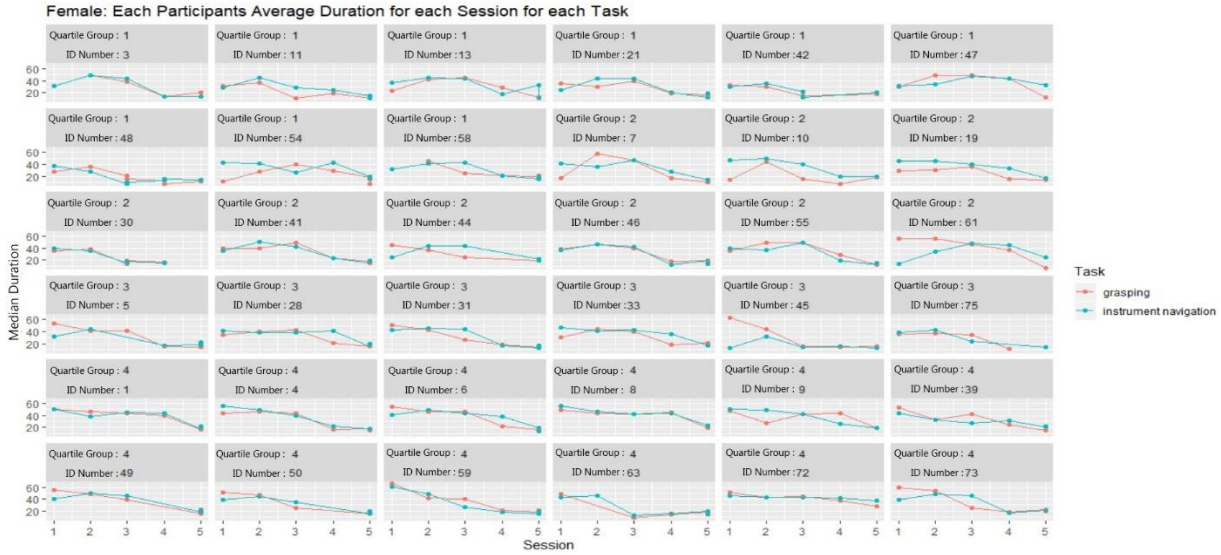
The overall aim of this section is to determine, using visual inspection, how many participants do not follow the population effect. These participants may be possible outliers and may affect the model and how it fits the data. On visual inspection, most individuals appear to follow the general population effect for both duration and damage count.

The population level effect for duration shows a decline at session 4, that continues until session 5, as seen from Figure D1. This indicates at the 4th session, the *inflection point* is reached where it takes more repetitions to increase learning performance, and therefore further practice does not result in significant performance increases (Pusic et al., 2015).

Only 22% [1] of female participants do not follow the population effect and do not show a decline for either task at session 4, as seen from Figure D3. At an individual level, about half of these female participants had an initial disadvantage and were also part of the low initial performance group. There were no substantial gender differences; 28%[2] of male participants did not show a decline in session 4, as seen from Figure D4. By session 5, 95%[3] of all participants showed a duration score improvement.

Figure D3

Individual Duration Learning Curves for Female Participants

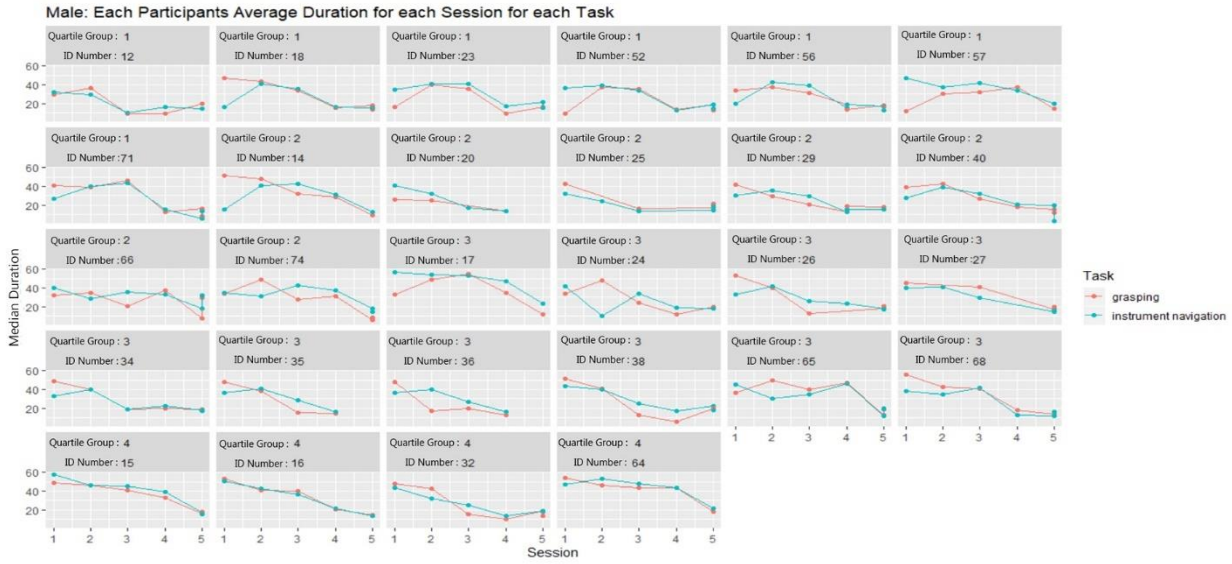


Note. The learning curve was made using the median score of the session for each participant. Each participant has two numbers in their grid. The top number indicates which quartile group (1,2,3 or 4) they belonged to; 1 is the high initial performing group, and 4 is the low initial performing group. The bottom number indicates each individual's identification number (ID).

[1] Participants 47, 44 61, 28, 1, 8, 49, 72
[2] Participants 57, 14, 66, 74, 17, 65, 15, 64
[3] Participants 66, 47 and 72 did not show a decline for one of the tasks

Figure D4

Individual Duration Learning Curves for Male Participants



Note. The learning curve was made using the median score of the session for each participant. Each participant has two numbers in their grid. The top number indicates which quartile group (1,2,3 or 4) they belonged to; 1 is the high initial performing group, and 4 is the low initial performing group. The bottom number indicates each individual's identification number (ID).

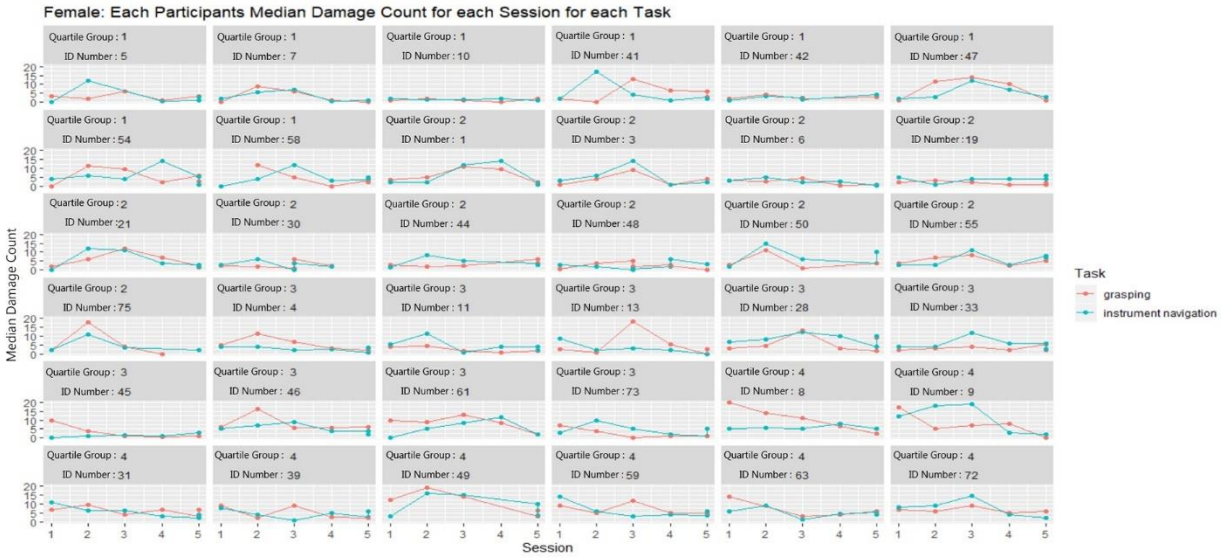
For damage count, as seen in duration, most participants follow the population level effect and again there were no gender differences. The population effect showed damage count had an increase from session 1 to 2, and even at session 3 there was poor performance, as seen in Figure D2.

This performance outcome was also seen at an individual level with 98%[4] of participants showing an accuracy decline at some point, as seen from visual inspection of Figures D5 and D6. Those that did not have a decline, showed they had already peaked optimal performance early in the training and continued to keep this accuracy throughout the sessions.

[4] Participants 10, 20, and 35 did not show an accuracy decline at some point as they already displayed optimal performance early in the training

Figure D5

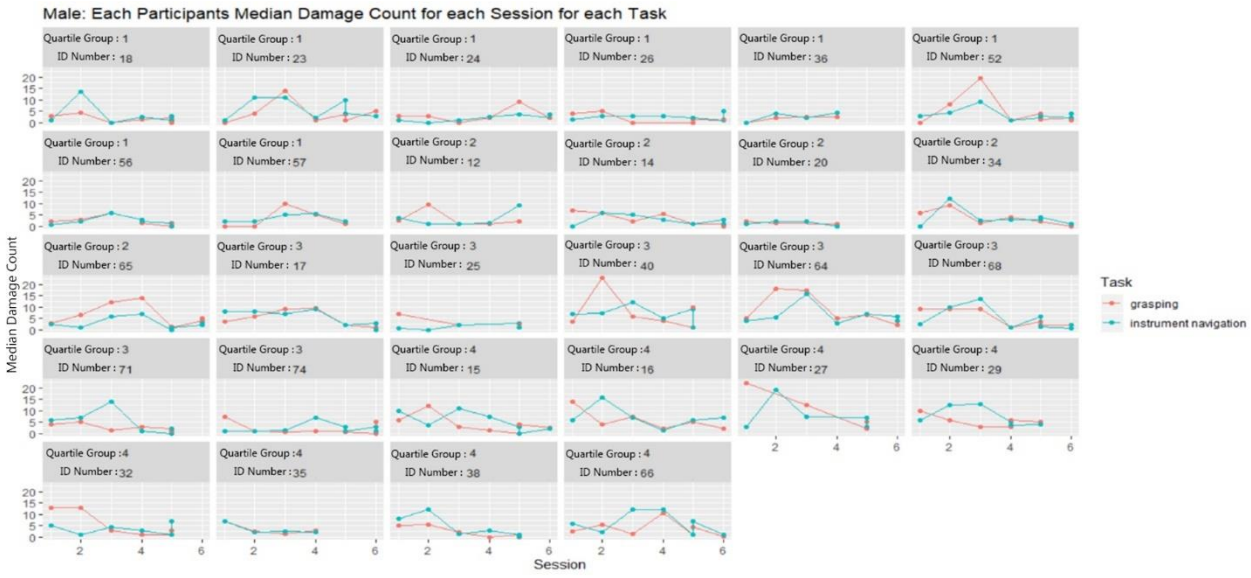
Individual Damage Count Learning Curves for Female Participants



Note. The learning curve was made using the median score of the session for each participant. Each participant has two numbers in their grid. The top number indicates which quartile group (1,2,3 or 4) they belonged to; 1 is the high initial performing group, and 4 is the low initial performing group. The bottom number indicates each individual's identification number (ID).

Figure D6

Individual Damage Count Learning Curves for Male Participants



Note. The learning curve was made using the median score of the session for each participant. Each participant has two numbers in their grid. The top number indicates which quartile group (1,2,3 or 4) they belonged to; 1 is the high initial performing group, and 4 is the low initial performing group. The bottom number indicates each individual's identification number (ID).

Appendix E

This section shows the results for individual noise within the data. It was found that about 40% of the noise in duration and damage count is produced by the individuals (Table E1 & Table E2). Therefore, both performance measures were equal in how much individual differences accounted for the noise and variation. Table E3 and Table E4 show which individual participants created the noise found in duration and damage count.

Table E1

Noise Caused by Individual Differences for Duration Performance

parameter	type	fixed effect	re_factor	center	lower	upper
Sigma	disp	NA	NA	13.74	13.33	14.17
Sigma Individual Level	grpef	Intercept	ID	5.12	4.12	6.37

Note. Individuals cause about a 40% (0.37 = 5.12/13.74) of the noise in duration.

Table E2

Noise Caused by Individual Differences for Damage Count Performance

parameter	type	fixed effect	re_factor	center	lower	upper
Sigma	disp	NA	NA	4.68	4.54	4.83
Sigma Individual Level	grpef	Intercept	ID	2.02	1.65	2.50

Note. Individuals cause about 40% (0.43 = 2.01 / 4.68) of the noise in damage count.

Table E3

Individual Differences in Duration

Identification Number (ID)	**center**	**lower**	**upper**
1	9.02	4.66	13.29
2	-0.55	-4.95	4.02
3	-3.17	-7.98	1.70
4	1.97	-2.78	6.82
5	1.52	-3.62	6.43
6	4.30	-0.06	8.84
7	1.43	-3.82	6.83
8	9.17	3.90	14.49
9	6.07	1.01	11.24
10	2.38	-2.04	6.91
11	-6.48	-11.38	-1.81
12	-6.73	-11.49	-1.83
13	-3.17	-8.16	1.68
14	-1.26	-4.87	2.51

15	6.71	2.48	11.03
16	4.69	-0.20	9.42
17	3.73	-0.95	8.47
18	-2.46	-7.96	3.03
19	-2.69	-6.40	0.99
20	-4.10	-9.31	1.08
21	-2.27	-6.68	1.98
23	-0.53	-5.38	4.30
24	-4.68	-10.05	0.73
25	-4.65	-10.00	0.30
26	-2.73	-7.63	2.02
27	2.24	-3.13	7.97
28	3.14	-1.28	7.44
29	-3.92	-8.05	0.32
30	-3.14	-8.00	1.38
31	2.69	-2.46	7.86
32	-5.46	-9.90	-1.09
33	1.60	-3.10	6.16
34	-0.05	-5.16	5.30
35	-5.43	-10.55	-0.31
36	-2.70	-8.14	2.49
38	-1.24	-5.81	3.46
39	-0.53	-4.90	3.76
40	-2.31	-7.12	2.73
41	-1.52	-6.55	3.38
42	-2.15	-6.28	2.10
44	1.54	-3.15	6.38
45	-7.18	-11.82	-2.54
46	3.58	-0.91	8.00
47	8.65	3.70	13.72
48	-7.46	-11.79	-3.16
49	8.34	3.75	12.72
50	3.62	-1.55	8.82
52	-6.39	-10.88	-1.75
54	-2.23	-6.56	1.91
55	-3.03	-7.14	1.38
56	-4.97	-9.09	-0.96
57	2.47	-2.51	7.27
58	-0.77	-5.58	4.27
59	1.14	-2.90	5.23
60	-4.76	-8.55	-0.95
61	7.93	2.86	13.08
63	-3.67	-8.13	0.89
64	9.55	4.24	14.65
65	3.45	-1.06	7.91
66	1.00	-3.42	5.50
67	-0.42	-5.09	4.37
68	-4.38	-8.62	-0.26

71	-5.26	-9.46	-1.06
72	8.06	2.61	13.72
73	3.46	-1.36	8.43
74	-0.04	-4.28	4.05
75	1.25	-4.66	7.07

Note. This table indicates how much each individual participant is off from the population level average, and their predicted response. The more off the center values if from 0 the more atypical the individual was.

Table E4

Individual Difference for Damage Count

Identification number (ID)	**center**	**lower**	**upper**
1	2.78	1.20	4.33
2	-1.24	-2.80	0.40
3	-0.24	-1.90	1.47
4	-0.06	-1.72	1.60
5	0.13	-1.69	1.88
6	-2.15	-3.77	-0.45
7	-1.18	-3.09	0.64
8	2.41	0.61	4.24
9	3.20	1.38	5.09
10	-2.83	-4.42	-1.33
11	-1.19	-2.84	0.45
12	-1.18	-2.84	0.42
13	-0.80	-2.51	0.90
14	-0.77	-2.15	0.60
15	-0.07	-1.67	1.45
16	1.94	0.34	3.57
17	0.38	-1.22	2.01
18	-1.28	-3.25	0.66
19	-1.69	-2.96	-0.39
20	-2.39	-4.28	-0.55
21	1.17	-0.35	2.71
23	1.41	-0.36	3.16
24	-2.06	-3.97	-0.15
25	-2.14	-4.04	-0.28
26	-2.44	-4.16	-0.77
27	4.24	2.26	6.24
28	1.50	-0.06	2.95
29	1.89	0.45	3.39
30	-1.36	-3.18	0.33
31	1.94	0.11	3.74
32	-0.56	-2.15	1.03
33	-0.36	-2.00	1.24
34	0.22	-1.68	2.21
35	-1.47	-3.36	0.33

36	-1.45	-3.36	0.41
38	-0.48	-2.11	1.14
39	-0.81	-2.32	0.72
40	1.78	0.09	3.48
41	-1.19	-2.92	0.51
42	-1.71	-3.23	-0.25
44	-0.42	-2.27	1.32
45	-2.24	-3.89	-0.61
46	1.50	-0.12	3.09
47	2.40	0.65	4.13
48	-2.04	-3.65	-0.45
49	6.00	4.36	7.71
50	1.69	-0.23	3.60
52	-0.64	-2.22	0.97
54	0.40	-1.10	1.90
55	0.21	-1.22	1.69
56	-2.05	-3.48	-0.70
57	0.01	-1.73	1.77
58	0.40	-1.38	2.17
59	0.77	-0.66	2.17
60	-2.71	-4.10	-1.33
61	2.85	1.02	4.64
63	0.73	-0.90	2.32
64	2.49	0.62	4.41
65	-0.03	-1.63	1.64
66	2.16	0.56	3.72
67	-1.12	-2.82	0.62
68	-0.57	-2.10	0.94
71	-0.46	-1.95	1.02
72	2.19	0.19	4.23
73	-0.64	-2.31	1.02
74	-2.25	-3.72	-0.74
75	-0.50	-2.62	1.69

Note. This table indicates how much each individual participant is off from the population level average, and their predicted response. The more off the center values if from 0 the more atypical the individual was.

Appendix F

Model criticism was performed in order to determine whether the prior distribution and model used correctly interpreted the raw data, and if the models themselves can be trusted.

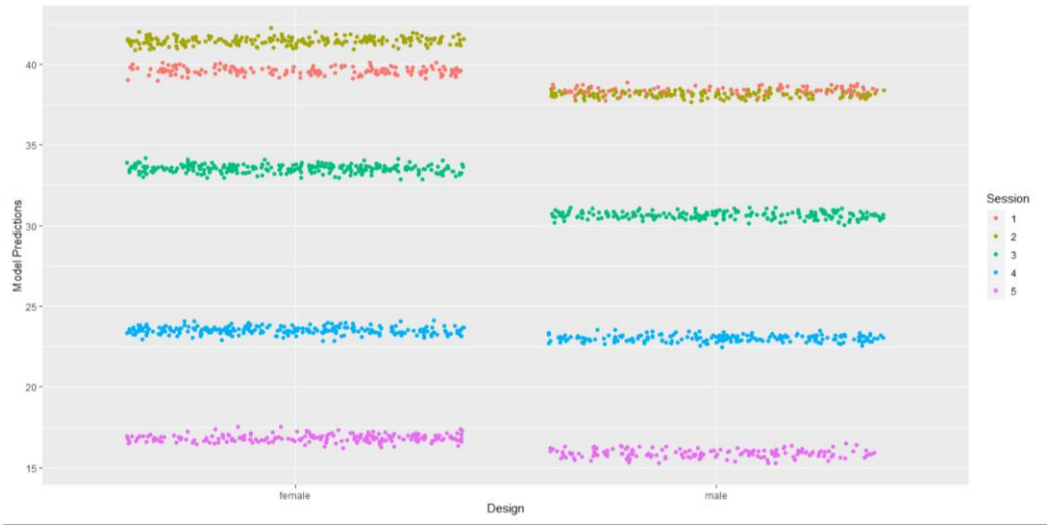
Gender Models

The multilevel models that split male and female participants were analysed.

Predicted Results. For duration, the sessions take account for why there are separate predicted groupings for the predicted scores (Figure F1). Furthermore, as the separate predicted groupings were gathered in close clusters rather than dispersed, this indicates that the predictions were not highly distributed and that it was quite similar within the groups. From these observations, it is possible to conclude that the model was not affected by predictions with different variations.

Figure F1

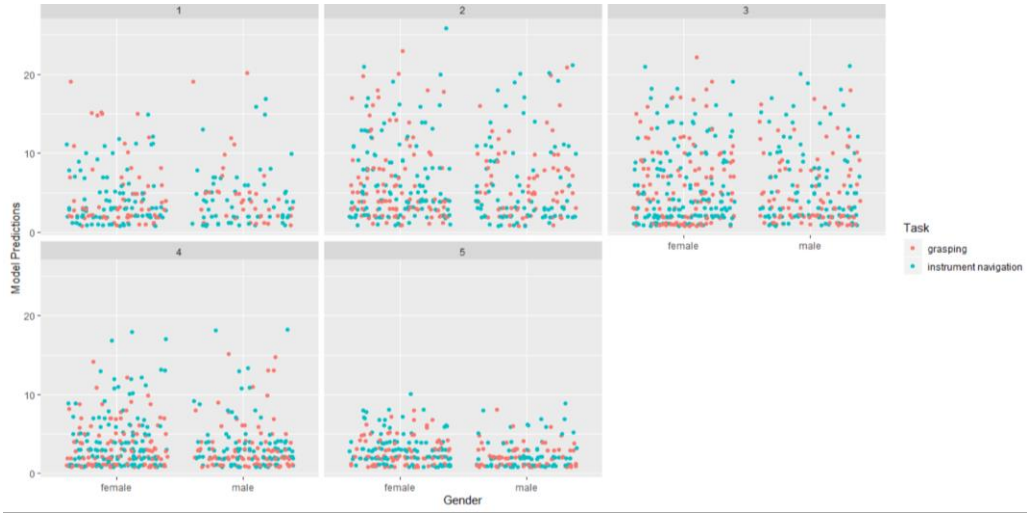
Gender Predictions for Duration, split into Sessions



For damage count scores there was a large variation. However, with practice, all participants regardless of gender, became less varied and showed similar performance outcomes, as seen in Figure F2. Additionally, both gender groups showed similar patterns with comparable prediction scores in terms of how distributed they were.

Figure F2

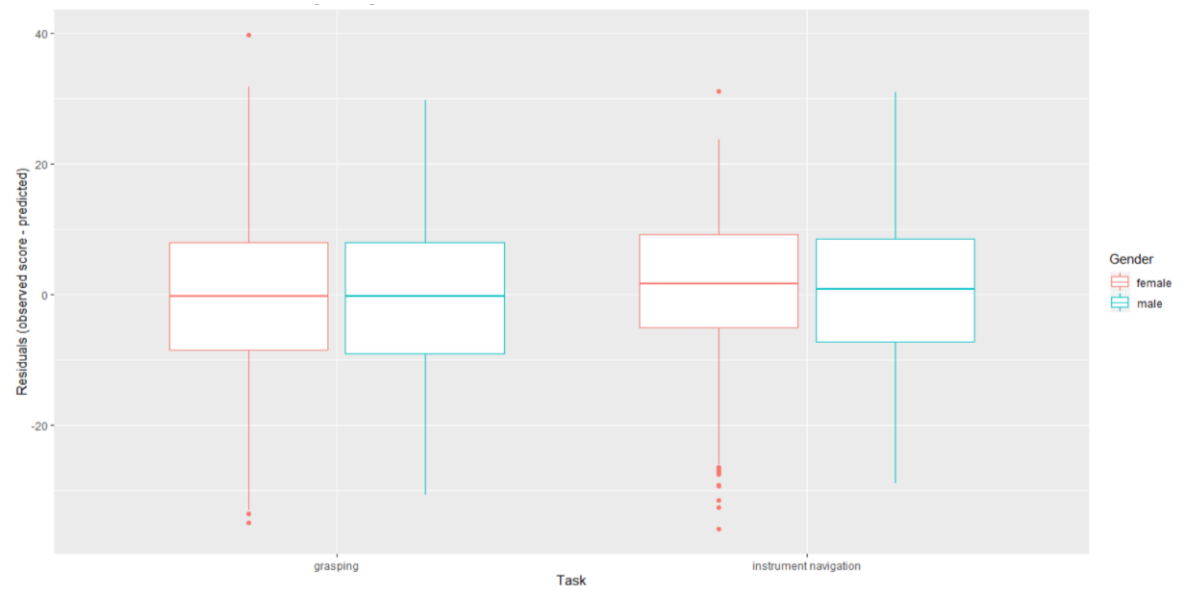
Gender Predictions for Damage Count, split into Sessions



Residual Analysis. This is the observed score minus the predicted score outputted from the model. For duration and damage count, there is not a large residual difference between the male and female groups (see Figure F3 & F4). Therefore, we can trust the models with a gender level did not have false results caused by range differences.

Figure F3

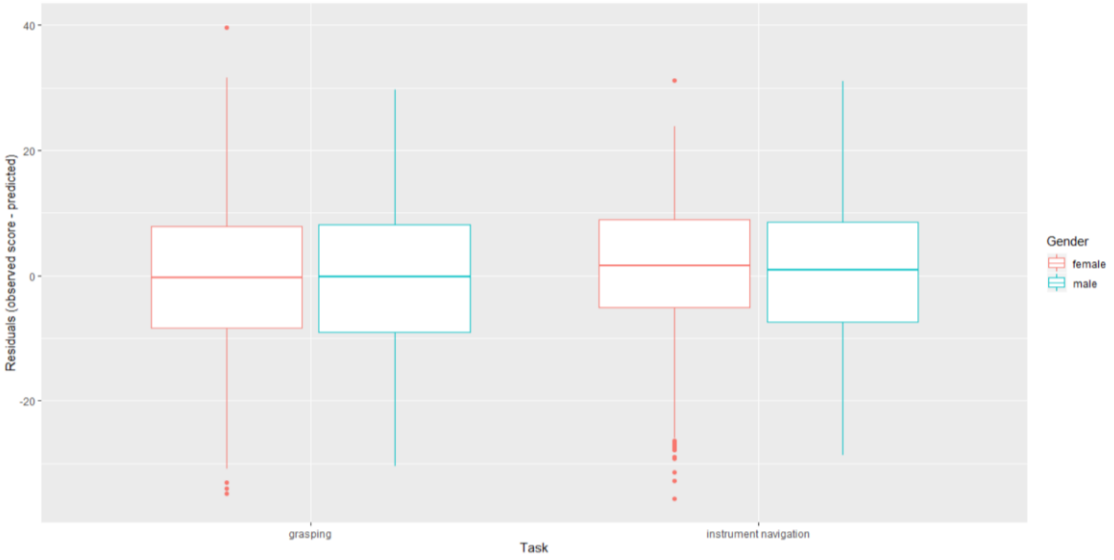
Duration Residuals for Gender Groups, using a Boxplot



Note. Actual duration scores in the experiment were compared to predicted duration scores to obtain standard deviation (residual). This is based on the model that used an ExGaussian Distribution.

Figure F4

Damage Count Residuals for Gender Groups, using a Boxplot



Note. Actual damage count scores in the experiment were compared to predicted damage count scores to obtain standard deviation (residual). This is based on the model that used a Poisson Distribution.

Predictive Power. It appears there is no predictive power of separating gender for duration or for damage count. For duration in Table F1, when comparing the predictive distribution that does not have gender as a random effect, there is no substantial difference and gender does not show a group effect (random factor variation). For example, for the intercept, the female group in session 1 took 39.63 seconds while the male group took 38.36 seconds (39.63 + - 1.27). When this particular model is compared with a model which that does not take account of gender, then in session 1 a participant will probably take 39.15 seconds [37.68,40.53]CI95% to complete the task. Both 39.63 seconds, the female group mean duration for session 1, and 38.36 seconds for the male group fit within the credibility interval whereby gender is not a contributing factor ([37.68,40.53]CI95%).

Table F1

Predictive Power of Duration for Gender; Comparing Model with Gender Groupings and a Model Without

Model	Fixed Effect	center	lower	upper
Model with Gender Groups	Intercept	39.63	37.84	41.33
	Session 2	1.84	-0.51	4.17
	Session 3	-6.05	-8.26	-3.78
	Session 4	-16.17	-18.31	-13.79
	Session 5	-22.82	-25.20	-20.40
	Male Group	-1.27	-4.17	1.60
	Session 2	-2.04	-5.71	1.61
	Session 3	-1.68	-5.20	1.75
	Session 4	0.81	-2.83	4.31
	Session 5	0.38	-3.51	4.19
Model without Gender Groups	Intercept	39.15	37.68	40.53
	Session 2	0.91	-0.94	2.77
	Session 3	-6.83	-8.63	-4.98
	Session 4	-15.84	-17.67	-14.04
	Session 5	-22.71	-24.57	-20.74

Note. This is the fixed effect output of the posterior distribution for two models. The distribution is the predicted values conditional on observed values (Schmettow, 2018).

As practice took place, the duration model showed it was unable to be a predictive variable to determine differences with training. For example, in reference to Table F1, from session 1 to session 5, the female group decreased their mean duration by 22.82 seconds and obtained a mean time of 16.81 seconds for session 5 (39.63 + -22.82). The male group decreased the mean duration by 23.71 seconds (-22.82 + -1.27 + 0.38), obtaining a mean time of 15.92 seconds for session 5 (39.63 + - 23.71). The decrease in mean duration from session 1 to session 5, for both the male and female groups (22.82 and 23.71 seconds, respectively), fits within the credibility interval for the model where gender was not taken into consideration [-24.57, -20.74]CI95%. This same pattern is seen for all the other sessions as well[5].

For damage count in Table F2, the intercept for the Poisson model for damage count (1.20[1.06,1.36]CI95%) was similar to the predicted intercept when using the Poisson model to compare gender groups for damage count (1.22 [1.03, 1.42]CI95%). As both values are within the

[5] Session 2, 3, and 4

credibility interval of the other model's intercept, we can assume that separating gender did not account for changes in damage count.

For the damage count model, there was also a minimal effect on the standard error. Therefore, splitting the groups by gender did not reduce the unknown error. This is apparent from Table F2 as the sigma for the model with a gender group and the model without are similar (0.41 and 0.43, respectively) and the credibility intervals have a large amount of overlap ([0.29, 0.56]CI95% and [0.32, 0.56]CI95%, respectively).

Table F2

Predictive Power of Damage Count for Gender; Comparing Model with Gender Groupings and a Model Without

Model	Type	Parameter	center	lower	upper
Model with Gender Groups	Fixed effect	Intercept	1.22	1.03	1.42
		Session 2	0.43	0.21	0.62
		Session 3	0.32	0.07	0.57
		Session 4	-0.21	-0.48	0.07
		Session 5	-0.31	-0.54	-0.09
		Male Group	-0.01	-0.33	0.28
		Session 2	-0.10	-0.43	0.23
		Session 3	-0.17	-0.58	0.21
		Session 4	-0.02	-0.42	0.40
		Session 5	-0.22	-0.56	0.15
	Sigma (Variation from Posterior Distribution)	Intercept	0.41	0.29	0.56
		Session 2	0.29	0.12	0.48
		Session 3	0.57	0.43	0.72
		Session 4	0.56	0.42	0.73
		Session 5	0.37	0.17	0.56
		Male Group	0.24	0.05	0.51
Model without Gender Groups	Fixed effect	Intercept	1.21	1.06	1.36
		Session 2	0.39	0.23	0.55
		Session 3	0.26	0.06	0.44
		Session 4	-0.21	-0.41	-0.01
		Session 5	-0.39	-0.59	-0.20
	Sigma (Variation from Posterior Distribution)	Intercept	0.43	0.32	0.56
		Session 2	0.27	0.10	0.49
		Session 3	0.56	0.42	0.73
		Session 4	0.53	0.39	0.70
		Session 5	0.42	0.24	0.61

Note. This is the fixed effect output of the posterior distribution for two models. The distribution is the predicted values conditional on observed values (Schmettow, 2018). A higher sigma value indicates more individual noise.

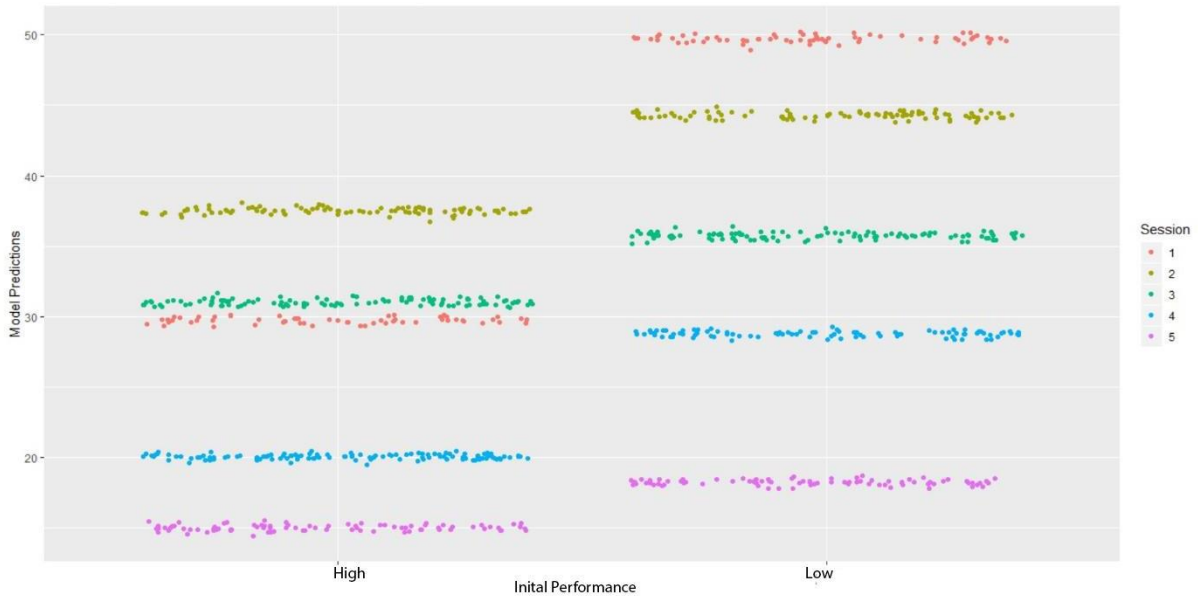
Initial Performance Models

The multilevel models that split high initial performers and low initial performers were analysed.

Predicted Results. For duration, the sessions take account for why there are separate predicted groupings for the predicted scores (Figure F5). Furthermore, as the separate predicted groupings were gathered in close clusters rather than dispersed, this indicates that the predictions were not highly distributed and that it was quite similar within the groups. These observations suggest that the model was not affected by predictions with different variations. The high initial performing group showed they had a performance decline with training. In Figure F5, they were predicted faster duration scores for the 1st session compared to sessions 2 and 3.

Figure F5

initial Performance Predictions for Duration split into sessions

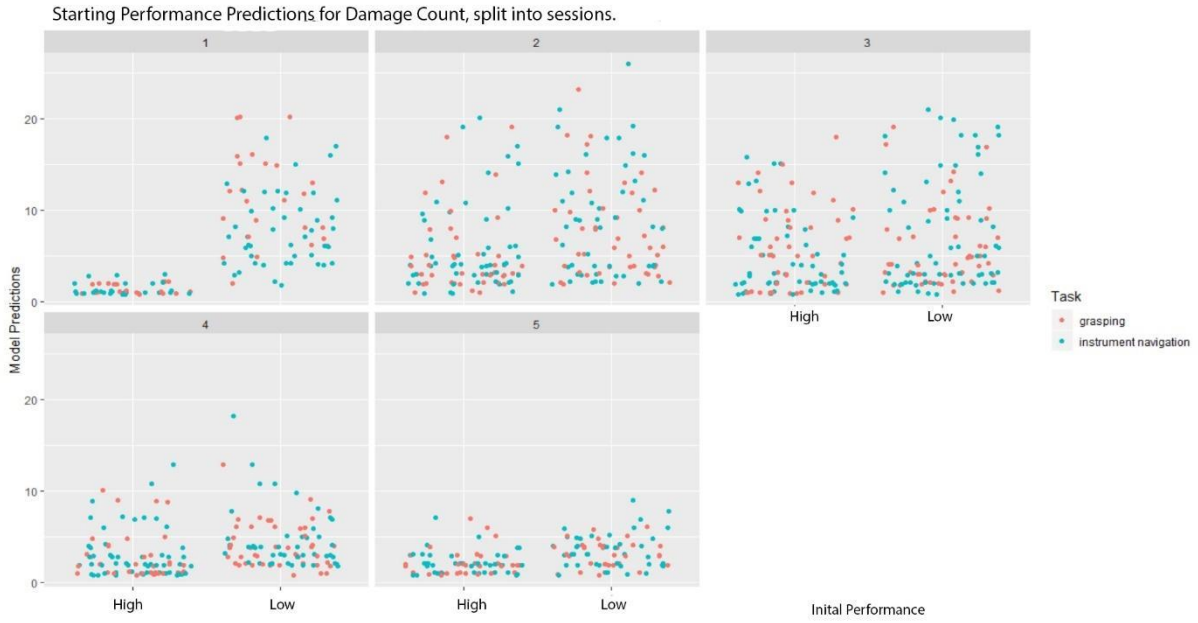


For damage count, it appeared that for the high initial performing group the session number influenced the distribution of the predicted scores. The pattern is strange as it indicates that with

practice individuals in this group had more varied predicted responses in the model. In Figure F6, for session 1, the high initial performing group acts in a much more uniform fashion compared to the low initial performing group. Furthermore, the high initial performing group only have these narrowly distributed predicted scores for the first session; for later sessions, the predicted scores are more spread out and show a similar pattern to the low initial performing group. Both groups have similar predicted distributions for sessions 2, 3, 4 and 5.

Figure F6

Initial Performance Predictions for Damage Count, Split into Sessions



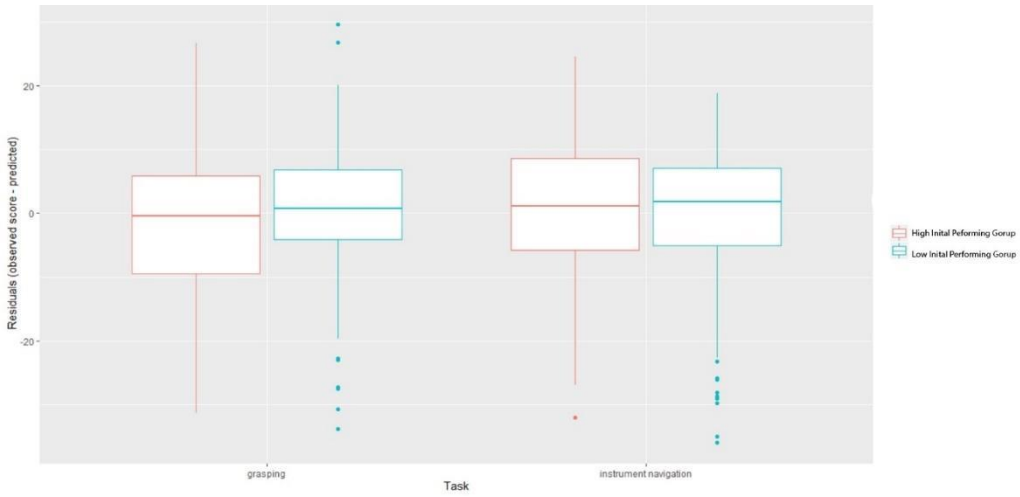
Residual Analysis. This is the observed score minus the predicted score outputted from the model. For both damage count and duration, the outcome seen by the initial performance models could have been produced because of the different residual ranges. Therefore, differences in initial performance groups and task may have led to false conclusions caused by range differences.

For duration, in Figure F7, the low initial performing group have smaller residuals (standard deviation) compared to the high initial performing group. This indicates their performance varied less than that of the population mean. On the other hand, the high initial performing group were

more atypical compared to the population and varied more regarding their individual performance. The grasping task also had the biggest variation between the two groups. Nevertheless, there were similarities and the mean predicted scores were all very similar.

Figure F7

Duration Residuals for Initial Performance Groups, using a Boxplot

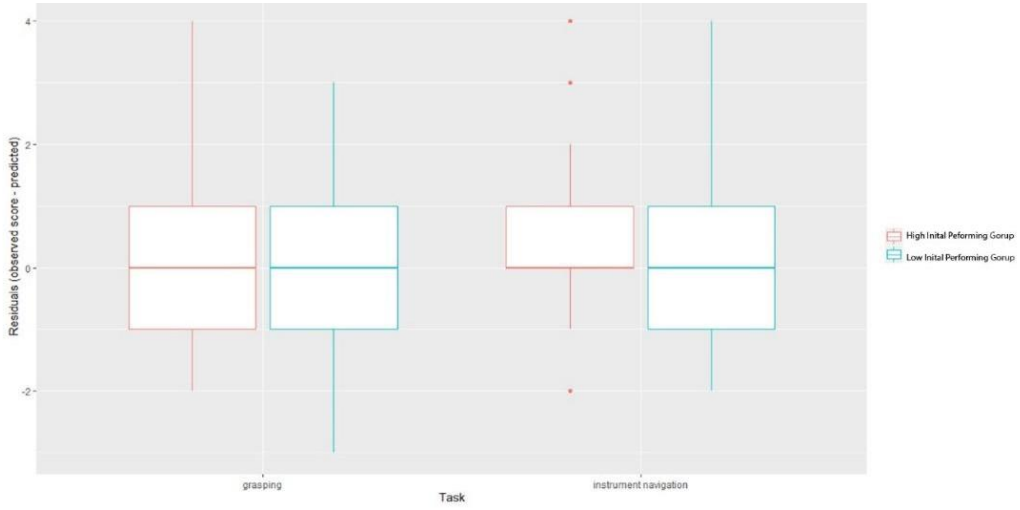


*Note***.** Actual duration scores in the experiment were compared to predicted duration scores to obtain standard deviation (residual). This is based on the model that used an ExGaussian Distribution.

For damage count, in Figure F8, the grasping task both initial performance groups have the same standard deviation (residual). For instrument navigation, the low initial performing group also has the same standard deviation as those for the gasping task. However, the high initial performing group is distinct and appear to have a smaller residual indicating they resemble the mean of the population to a greater extent. Therefore, the different groups had different residual ranges.

Figure F8

Damage Count Residuals for Initial Performance Groups, using a Boxplot



Note. Actual damage count scores in the experiment were compared to predicted damage count scores to obtain standard deviation (residual). This is based on the model that used a Poisson Distribution.

Predictive Power. The indications are that the initial performance model does hold predictive power, as suggested by the credibility intervals. This was seen for both duration and damage count models.

For duration, when referring to Table F3, the intercept for a model that does not take account of initial performance is 39.15 seconds [37.67,40.53]CI95%. When this is compared to the intercept of the high initial performing group, there was no overlap with the credibility intervals, therefore, it is certain that they were faster and took 29.76 seconds [26.98, 32.70]CI95%. From this it can be concluded that adding initial performance groups to the multilevel model created a group that, with 95% certainty, had different duration scores.

Table F3

Predictive Power of Duration for Initial Performance; Comparing Model with Initial Performance Groups and Model Without

Model	Fixed Effect	center	lower	upper
Model with Initial Performance Groups	Intercept	29.76	26.98	32.70
	Session 2	7.79	4.04	11.25
	Session 3	1.30	-2.28	4.87
	Session 4	-9.70	-13.20	-6.25
	Session 5	-14.75	-18.51	-11.21
	Low Performing Group	19.92	15.93	23.74
	Session 2	-13.18	-18.07	-7.96
	Session 3	-15.24	-20.14	-10.34
	Session 4	-11.18	-16.07	-6.38
	Session 5	-16.69	-21.66	-11.45
Model without Initial Performance groups	Intercept	39.15	37.68	40.53
	Session 2	0.91	-0.94	2.77
	Session 3	-6.83	-8.63	-4.98
	Session 4	-15.84	-17.67	-14.04
	Session 5	-22.71	-24.57	-20.74

Note. This is the fixed effect output of the posterior distribution for two models. The distribution is the predicted values conditional on observed values (Schmettow, 2018).

As practice took place, the initial performance duration model showed it was able to be a predictive variable to determine differences with training. For example, in reference to Table F3, from session 1 to session 2, the high initial performing group had an increase in speed of 7.79 seconds [4.04, 11.25], obtaining a mean time of 37.55 seconds for session 2 (29.76 + 7.79). The low initial performing group had an increase in speed of 14.53 seconds (7.79 + 19.92 + - 13.18), with a mean time of 44.29 seconds for session 2 (29.76 + 14.53). The increase in mean duration from session 1 to 2, for both the high initial performance group and the low initial performance group (7.79 and 14.53 seconds, respectively) does not fit within the credibility interval for the model where initial performance was not taken into consideration [-0.94, 2.77]CI95%. This indicates adding initial performance as a level, created groups, that with training, produced different changes in duration, compared to if no level had been added to the model. This was a pattern also seen for the other sessions[6].

[6] Sessions 3,4 and 5

For damage count, from Table F4, the intercept for the Poisson model for damage count was 1.21[1.06, 1.36]CI95% which is quite different to the predicted intercept when using the Poisson model to compare initial performance groups. In this model, the high initial performing group had an intercept of 0.33[0.01, 0.61]CI95%. As the credibility intervals are quite narrow and do not overlap, this indicates that having a multilevel model with a level for initial performance does change how damage count is predicted.

With regard to damage count, adding initial performance to the model did not necessarily change the standard error. Although, it appeared that splitting the groups resulted in less individual noise 0.16[0.04, 0.33]CI95%, compared to when only damage count was analysed 0.43[0.32, 0.56]CI95% (Table F4). This reduction in noise may have not necessarily occurred as there was slight overlap with the credibility intervals. Therefore, we cannot be entirely certain that the standard error was changed

Table F4

Predictive Power of Damage Count for Initial Performance; Comparing Model with Initial Performance Groups and Model Without

Model	Type	Parameter	center	lower	upper
Model with Initial Performance Groups	Fixed effect	Intercept	0.33	0.01	0.61
		Session 2	1.16	0.79	1.54
		Session 3	1.07	0.65	1.50
		Session 4	0.45	0.02	0.87
		Session 5	0.36	-0.04	0.76
		Low Performing Group	1.72	1.36	2.09
		Session 2	-1.32	-1.80	-0.83
		Session 3	-1.50	-2.07	-0.96
		Session 4	-1.17	-1.70	-0.61
		Session 5	-1.28	-1.80	-0.76
	Sigma (Variation from Posterior Distribution)	Intercept	0.16	0.04	0.33
		Session 2	0.35	0.18	0.55
		Session 3	0.51	0.35	0.72
		Session 4	0.48	0.29	0.70
		Session 5	0.30	0.10	0.56
		Low Performing Group	0.15	0.03	0.35
Model without Initial Performance Groups	Fixed Effect	Intercept	1.21	1.06	1.36
		Session 2	0.39	0.23	0.55
		Session 3	0.26	0.06	0.44
		Session 4	-0.21	-0.41	-0.01
		Session 5	-0.39	-0.59	-0.20
	Sigma (Variation from Posterior Distribution)	Intercept	0.43	0.32	0.56
		Session 2	0.27	0.10	0.49
		Session 3	0.56	0.42	0.73
		Session 4	0.53	0.39	0.70
		Session 5	0.42	0.24	0.61

Note. This is the fixed effect output of the posterior distribution for two models. The distribution is the predicted values conditional on observed values (Schmettow, 2018). A higher sigma value indicates more individual noise.

Task Type

It was found that the task either being grasping or instrumental navigation did not have a huge impact on the model criticism. As all the predicted results from all the multilevel models showed they were not affected by task type. All the figures below (Figure F9, F10, F11, & F12) plot the expected results produced by the multilevel models. On visual inspection, the type of task did not influence the expected results. Nevertheless, any task differences that were found have been indicated above.

Figure F9

Gender Predictions for Duration

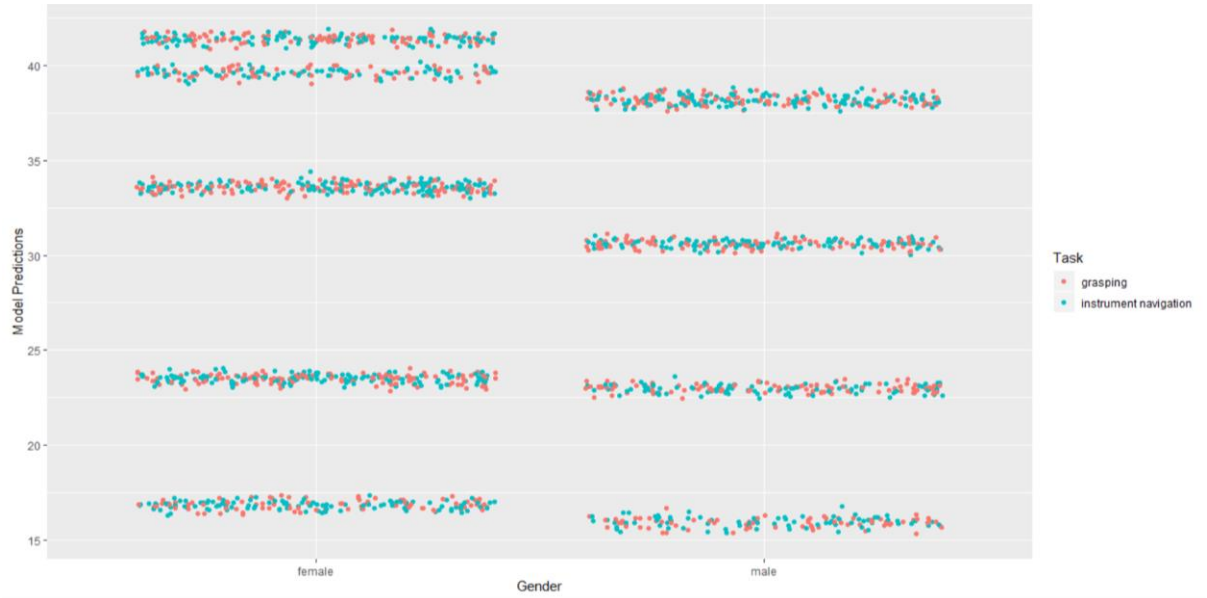


Figure F10

Gender Group Predictions for Damage Count

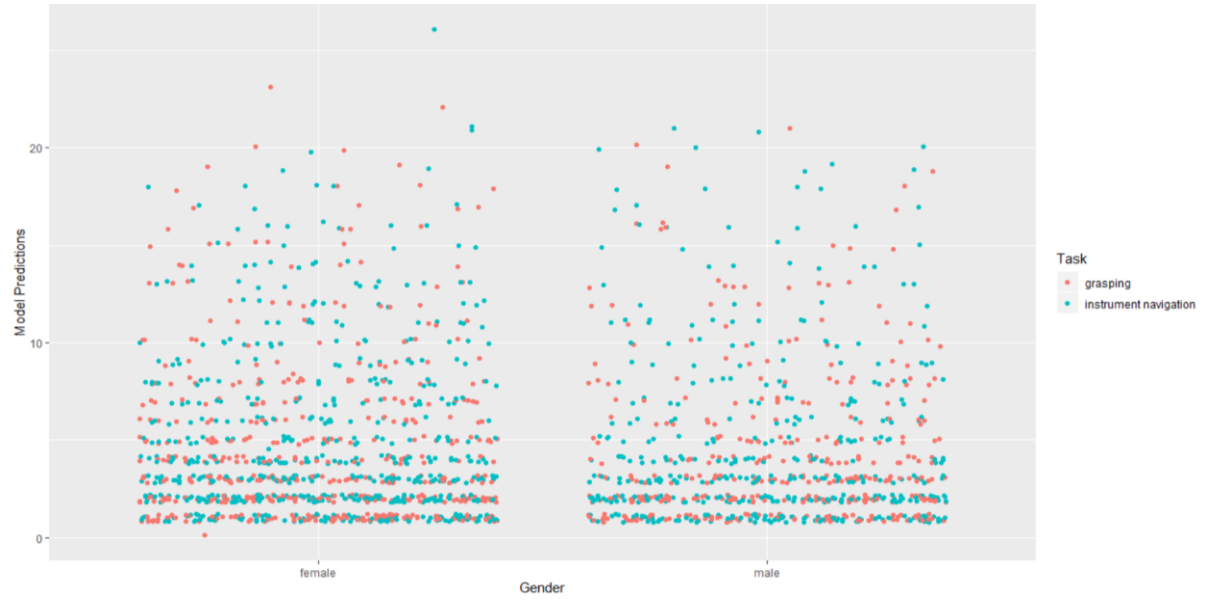


Figure F11

Initial Performance Predictions for Duration

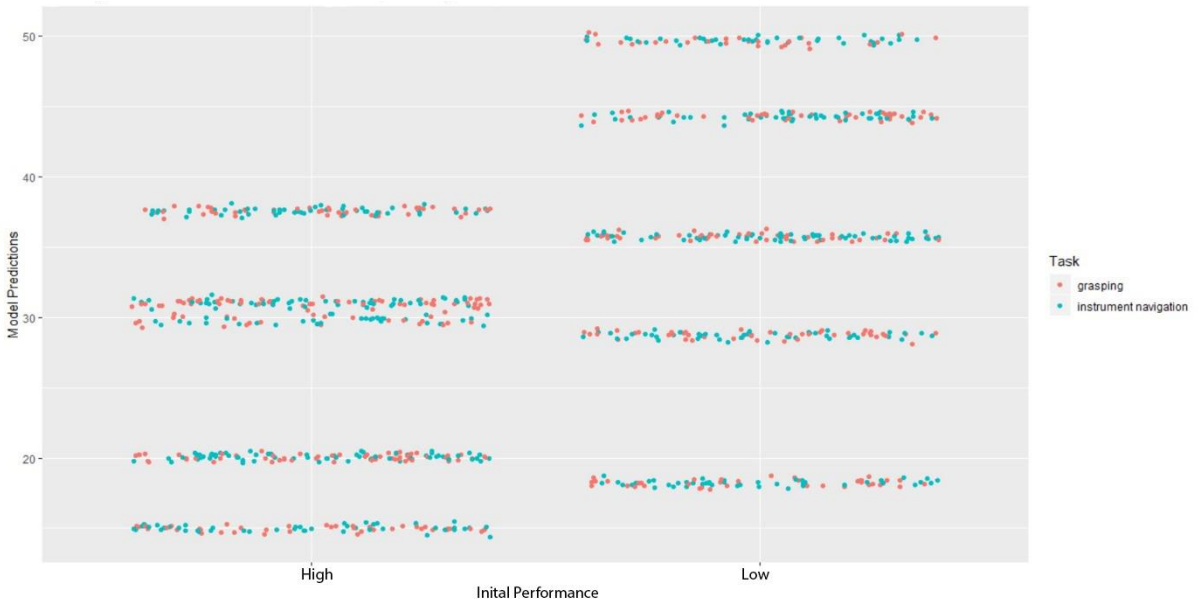
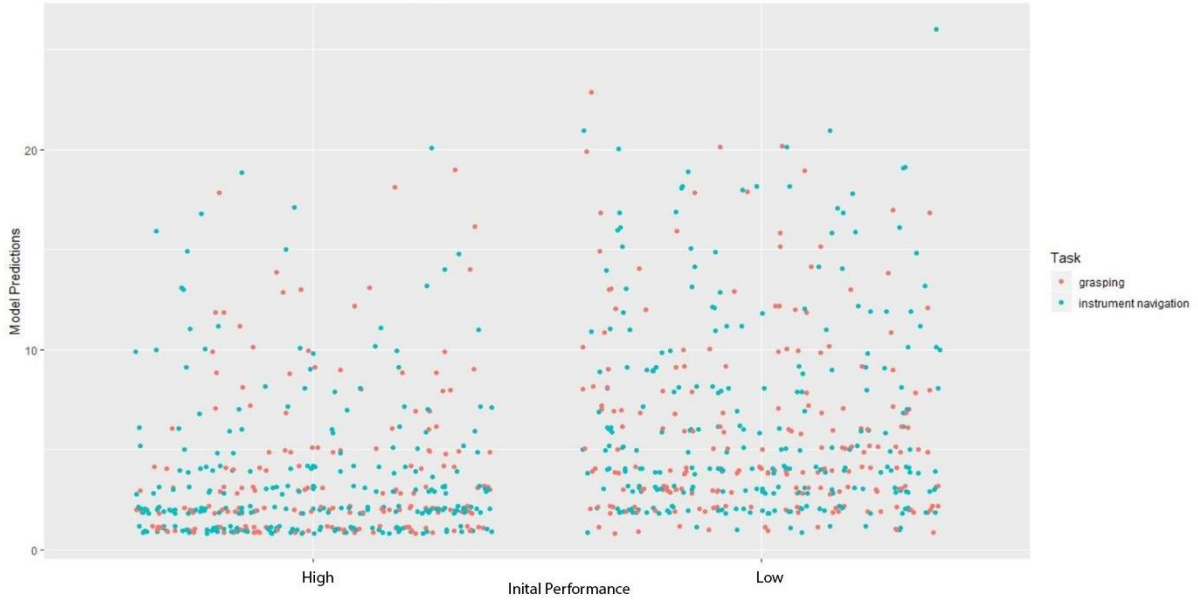


Figure F12

Initial Performance Predictions for Damage Count



Model Fit for Duration

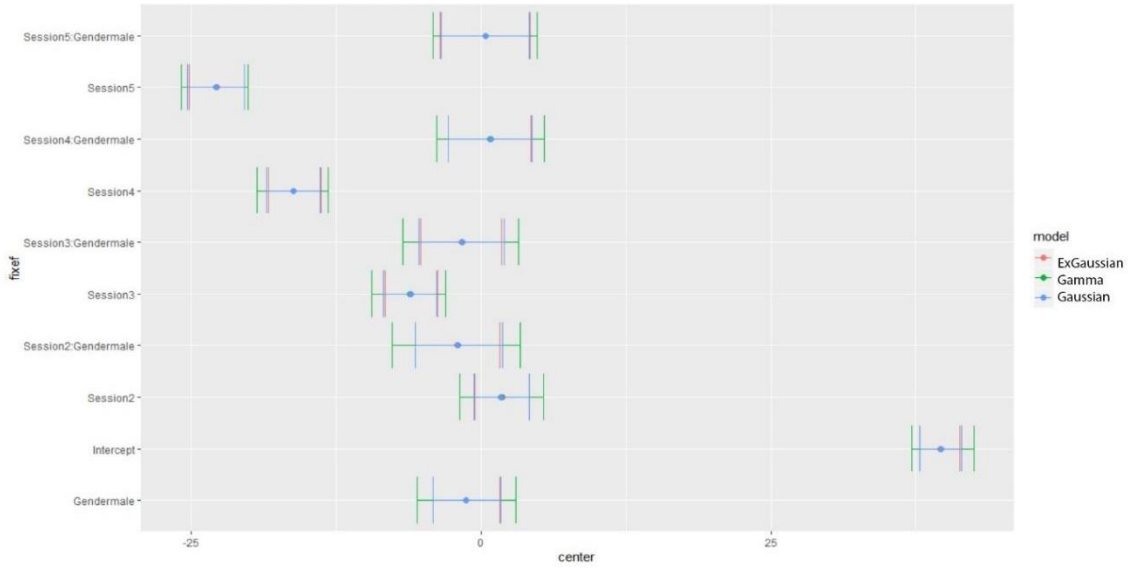
For duration data all the distributions, ExGaussian, Gaussian, and Gamma, do not hold a great fit in terms of residual analysis. Therefore, the bimodal nature of the data makes it hard to fit the data to an appropriate prior distribution. The use of an ExGaussian distribution was comparable

to a Gaussian distribution, and both these models have a better model fit than a Gamma distribution.

The Gamma distribution has the worst fit. When looking at fixed estimates in Figures F13 and F14, Gamma has the largest credibility estimates, compared to the Gaussian and ExGaussian distributions. The ExGaussian and Gaussian distributions are almost identical, shown by the similar gaps between the upper and lower bounds indicated by the vertical lines.

Figure F13

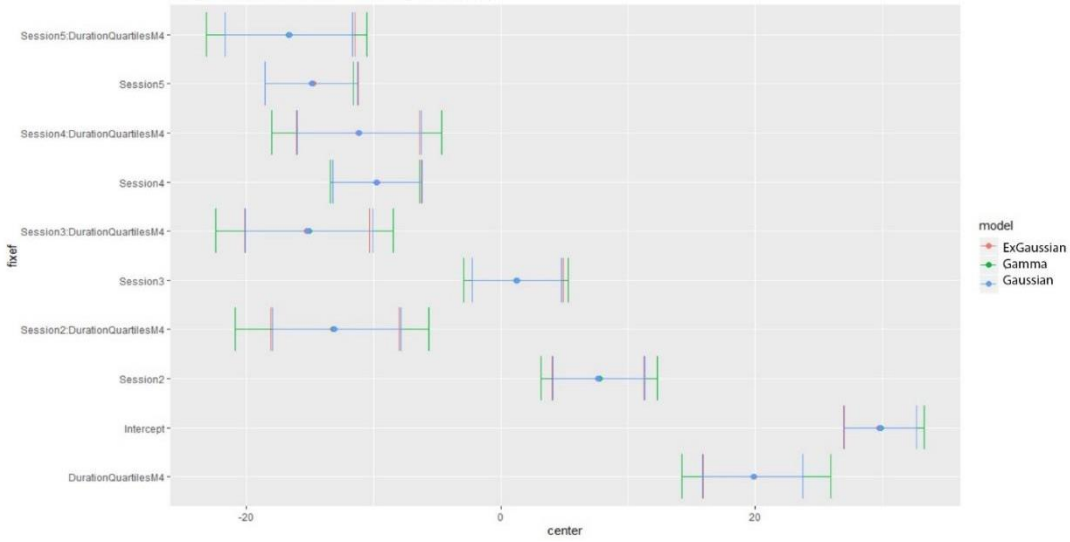
Gender Groups - Fixed Effect Estimates of Different Distributions



Note. Credibility intervals shown by the vertical lines, based on if duration data used either an ExGaussian, Gaussian, or Gamma prior distribution.

Figure F14

Initial Performance Groups - Fixed Effect Estimates of Different Distributions

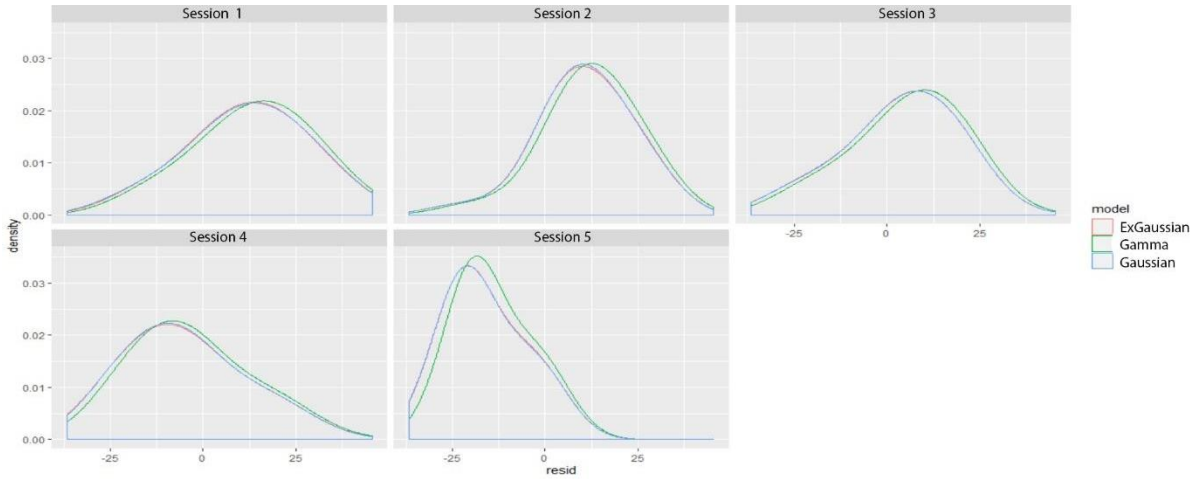


Note. Credibility intervals shown by the vertical lines, based on if duration data used either an ExGaussian, Gaussian, or Gamma prior distribution.

All distributions had a bad model fit as the peak (mode) of the residual distributions for all sessions are quite off from the target centre residual of 0, as seen in Figure F15 and F16. A good fit would have the modal peak of the distributions more around the centre (x-axis = 0).

Figure F15

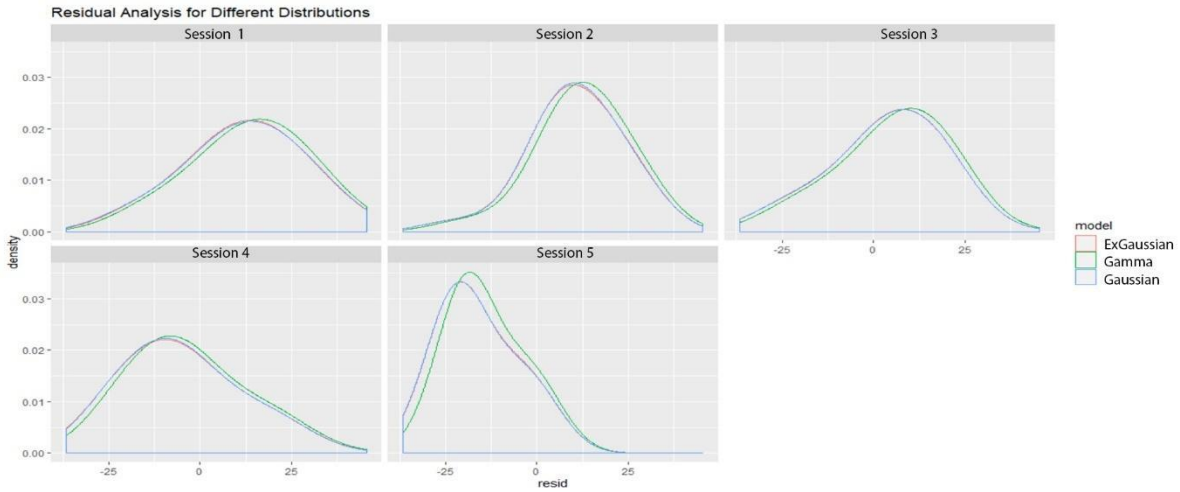
Gender Groups - Residual Analysis for Different Distributions, for each Session.



Note. Based on if the model for duration data used either an ExGaussian, Gaussian, or Gamma prior distribution. The residual centre is equal to 0 on the x-axis.

Figure F16

Initial Performance Groups - Residual Analysis for Different Distributions, for each Session



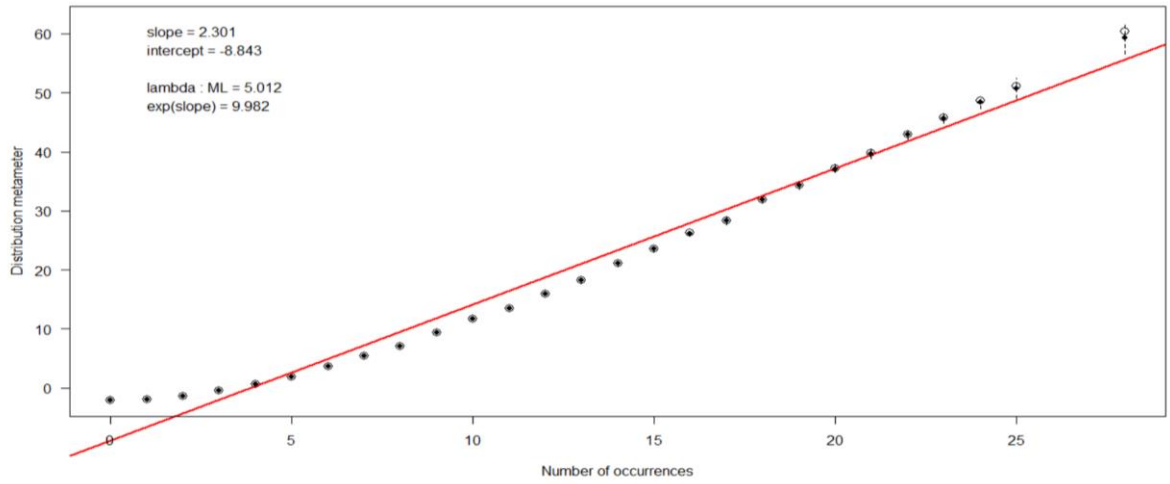
Note. Based on if the model for duration data used either an ExGaussian, Gaussian, or Gamma prior distribution. The residual centre is equal to 0 on the x-axis.

Model Fit for Damage Count

Q-Qplot were made by plotting observed frequency over the fitted frequency of the chosen distribution. The Poisson distribution fit the distribution of count data the best. For Figures F17, F18, and F19, the closer the observed frequency points are to the theoretical (red line), the better the distribution fits the data. The Poisson had the best fit (Figure F17), with negative binomial being a second runner up as although all the points fit the line quite well, the credibility intervals are quite wide for when there are a larger number of occurrences (Figure F18). Binomial did not fit the theoretical line and showed a curved shape indicating it is not a good distribution to represent the data (Figure F19).

Figure F17

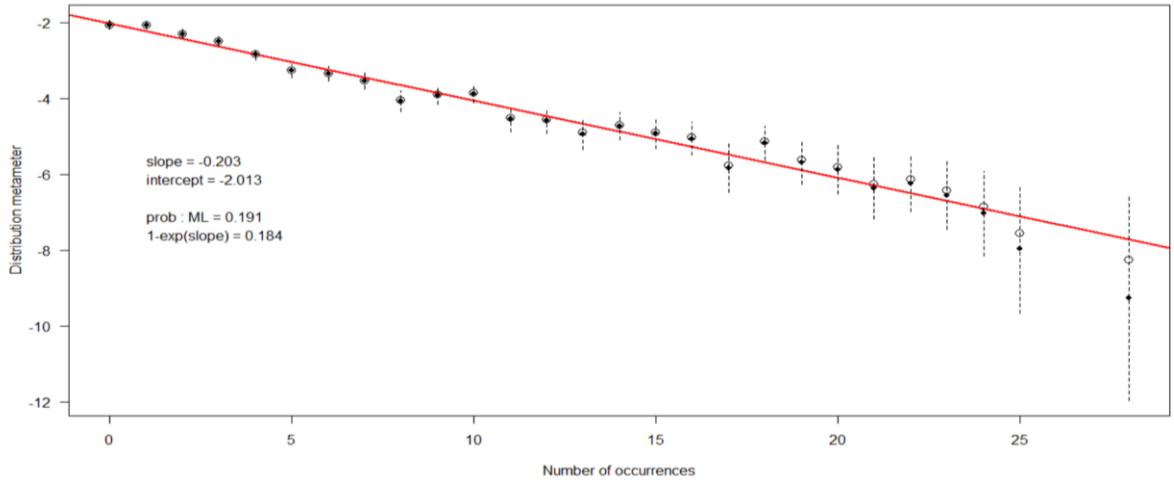
Q-Qplot showing Damage Count data placed with a Poisson Distribution.



Note. The closer the observed frequency points are to the theoretical (red line), the better the distribution fits the data. The vertical lines at each data point indicate credibility intervals; wider intervals indicate a poor model fit. A Poisson Distribution although not perfect at the ends does follow the theoretical line fairly closely.

Figure F18

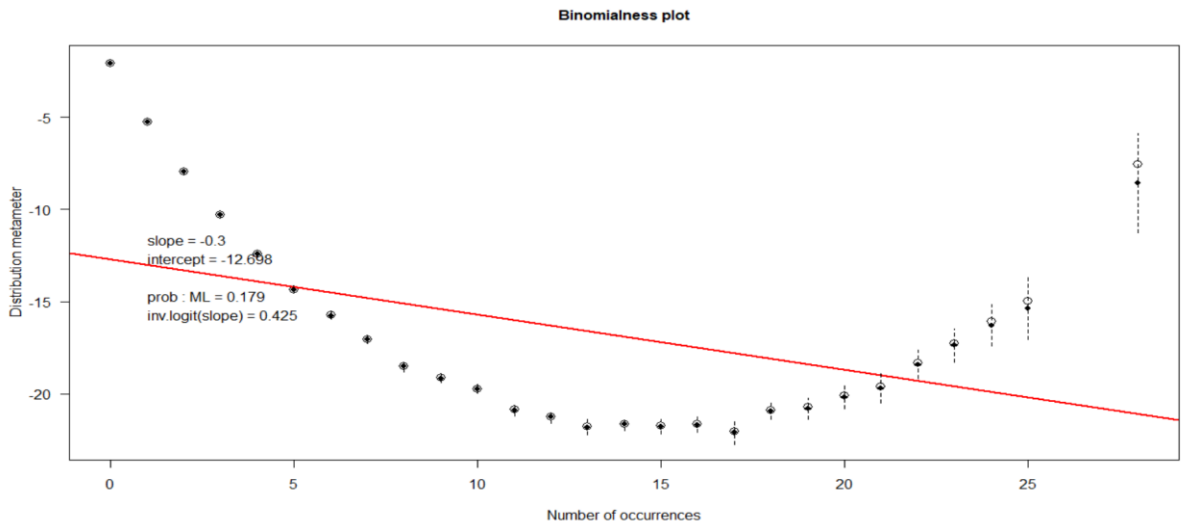
Q-Qplot showing Damage Count data placed with a Negative Binomial Distribution



Note. The closer the observed frequency points are to the theoretical (red line), the better the distribution fits the data. The vertical lines at each data point indicate credibility intervals; wider intervals indicate a poor model fit. The Gamma distribution shows wide credibility intervals when there were many occurrences.

Figure F19

Q-Qplot showing Damage Count data placed with a Binomial Distribution



Note. The closer the observed frequency points are to the theoretical (red line), the better the distribution fits the data. The vertical lines at each data point indicate credibility intervals; wider intervals indicate a poor model fit. The Binomial distribution does not follow the theoretical line indicating poor model fit.

Appendix G

Table G1

Output for ExGaussian model for Duration, taking account of Sessions and Gender Group

Fixed Effect	center	lower	upper
Intercept	39.63	37.84	41.33
Session 2	1.84	-0.51	4.17
Session 3	-6.05	-8.26	-3.78
Session 4	-16.17	-18.31	-13.79
Session 5	-22.82	-25.20	-20.40
Male Group	-1.27	-4.17	1.60
Session 2	-2.04	-5.71	1.61
Session 3	-1.68	-5.20	1.75
Session 4	0.81	-2.83	4.31
Session 5	0.38	-3.51	4.19

Note. The output was obtained using the *fixef* function and can be examined in a summative procedure.

Table G2

Calculations Needed to obtain Mean Score for the Gender Groups Duration Model

Session	Female Group Mean	Male Group Mean
Session 1	N/A	39.63 + -1.27
Session 2	39.63 + 1.84	39.63 + 1.84 + -1.27 + -2.04
Session 3	39.63 + -6.05	39.63 + -6.05 + -1.27 + -1.68
Session 4	39.63 + -16.17	39.63 + -16.17 + -1.27 + 0.81
Session 5	39.63 + -22.82	39.63 + -22.82 + -1.27 + 0.38

Table G3

Output for a Poisson model for Damage Count, taking account of Sessions and Gender Groups

Fixed Effect	center	lower	upper
Intercept	3.40	2.80	4.12
Session 2	1.53	1.24	1.86
Session 3	1.37	1.07	1.76
Session 4	0.81	0.62	1.07
Session 5	0.73	0.58	0.91
Male Group	0.99	0.72	1.33
Session 2	0.90	0.65	1.25
Session 3	0.84	0.56	1.24
Session 4	0.98	0.66	1.49
Session 5	0.80	0.57	1.17

Note. The output was obtained using the *fixef* and an exponential mean function and can be examined using a multiplicative procedure.

Table G4

Output for ExGaussian model for Duration, taking account of Sessions and Initial Performance Groups.

Fixed Effect	center	lower	upper
Intercept	29.76	26.98	32.70
Session 2	7.79	4.04	11.25
Session 3	1.30	-2.28	4.87
Session 4	-9.70	-13.20	-6.25
Session 5	-14.75	-18.51	-11.21
Low Performing Group	19.92	15.93	23.74
Session 2	-13.18	-18.07	-7.96
Session 3	-15.24	-20.14	-10.34
Session 4	-11.18	-16.07	-6.38
Session 5	-16.69	-21.66	-11.45

Note. The output was obtained using the *fixef* function and can be examined in a summative procedure.

Table G5

Calculations Needed to obtain Mean Score for Initial Performance Groups Duration Model

Session	High Initial Performing Group Mean	Low Initial Performing Group Mean
Session 1	N/A	29.76 + 19.92
Session 2	29.09 + 7.79	29.76 + 7.79 + 19.92 + -13.18
Session 3	29.76 + 1.30	29.76 + 1.30 + 19.92 + -15.24
Session 4	29.87 + -9.70	29.76 + -9.70 + 19.92 + -11.18
Session 5	29.76 + -14.75	29.76 + -14.75 + 19.92 + -16.69

Table G6

Output for a Poisson model for Damage Count, taking account of Sessions and Initial Performance Groups

Fixed Effect	center	lower	upper
Intercept	1.38	1.01	1.85
Session 2	3.18	2.20	4.68
Session 3	2.90	1.91	4.46
Session 4	1.56	1.02	2.39
Session 5	1.44	0.96	2.14
Low Initial Performing Group	5.60	3.91	8.05
Session 2	0.27	0.17	0.44
Session 3	0.22	0.13	0.38
Session 4	0.31	0.18	0.55
Session 5	0.28	0.17	0.47

Note. The output was obtained using the *fixef* and an exponential mean function and can be examined using a multiplicative procedure.