SAICSIT Conference 2021

Proceedings of the South African Institute of Computer Scientists and Information Technologists

September 2021

Organising Committee

Organizing chairs: Prof Shawren Singh and Dr Mathias Mujinga|
Programme chairs: Dr Jan Mentz, Prof Hugo Lotriet and Prof Bobby Tait
Conference secretary: Prof Bester Chimbo
M&D Symposium chair: Prof Kirstin Krauss (krauske@unisa.ac.za)
Panel chair: Prof Shawren Singh
Workshop chair: Prof Bester Chimbo
Keynote chair: Prof Steven Furnell (University of Nottingham)
Prof Judy van Biljon (Unisa)
Treasurer: Dr Mathias Mujinga
Information specialist: Dr Filistéa Naudé

UNISA SCHOOL OF COMPUTING &

UNIVERSITY OF SOUTH AFRICA

PRETORIA

https://Booksup.co.za/index.php/unisapress/catalog/book/127

https://doi.org/10.25159/127
ISBN 978-1-77615-121-9 (Online)

Print ISBN: 978-1-77615—120-2

E-ISBN: 978-1-77615-121-9

Attribution – Please cite the work as follows:

Singh, Shawren, Mujinga, Mathias, Lotriet, Hugo; Tait, Bobby and Mentz, Jan. 2021. *SAICSIT Conference 2021. Proceedings of the South African Institute of Computer Scientists and Information Technologists*. Pretoria: Unisa School of Computing.

Published by the Unisa School of Computing
with the University of South Africa Press, Pretoria.

Published October 2021.

Contents

List of Authors

We wish to thank the following for their contributions to this volume (in alphabetical order of surname):

Jecton Tocho Anyango, University of Cape Town, South Africa

Christine Asaju, University of the Witwatersrand, South Africa

Irwin Brown, University of Cape Town, South Africa

Mpumelelo Dhlamini, University of Cape Town, South Africa

Stephen Phillip Levitt, University of the Witwatersrand, South Africa

Geoffrey Lydall, University of the Witwatersrand, South Africa

Ken J. Nixon, University of the Witwatersrand, South Africa

Grant Oosterwyk, University of Cape Town, South Africa

Kiren Kosygin Padayachee, University of the Witwatersrand, South Africa

Hussein Suleman, University of Cape Town, South Africa

Hima Vadapalli, University of the Witwatersrand, South Africa

Message from the Conference Chairs

"If you're in a bad situation, don't worry it'll change. If you're in a good situation, don't worry it'll change."
John A Simone Sr

2021 has been an interesting year not only for academe but also for the academic conferencing community. Good academic ideas are not halted by any calamities, this year SAICSIT boldly held its second virtual conference. Naturally, a large number of academic researchers have been fatigued by the increase in online academic interactions. However, the University of South Africa, School of Computing saw an opportunity to actively host a conference for the SAICSIT community.

This hosting opportunity brought together a dedicated group of enthusiastic individuals who converted a survival conference into a reality. I am greatful to the following champions: Mathias Mujinga my co-chair, Jan Mentz, Hugo Lotriet, and Bobby Tait the programme chairs. Bester Chimbo the conference secretary. Kirstin Krauss the M&D Symposium chair and Filistéa Naudé. You sacrificed your Friday afternoons for a higher cause than you. I am further great full to Prof Steven Furnell and Prof Judy van Biljon for immediately agreeing to be keynote speakers. The College of Science, Engineering & Technology at Unisa through Prof Bhekie Mamba and the Head of Research & Graduate Studies, Prof SJ Johnston for the moral and financial support. Finally, at short notice Mr Pieter Rall and Mrs Hetta Pieterse from Unisa Press, who rearranged their production schedule to accommodate the SAICSIT proceedings.

In light of this year's theme "Reimagining the Interconnected World", this hosting opportunity has brought the SAICSIT community a little closer and has brought the organising committee much closer.

On behalf of the Organising Committee, we would like to thank all the reviewers and the conference participants of this virtual conference.

Conference Chair and Co-chair
Shawren Singh and Mathias Mujinga

Message from the Programme Chairs of SAICSIT 2021

On behalf of the Unisa School of Computing we would like to welcome you to the virtual SAICSIT 2021 conference.

A total of 30 submissions were received from authors in South Africa and abroad in response to two calls for papers. All papers were subjected to a stringent double-blind peer review process, with every paper reviewed by at least two reviewers. Ultimately 5 full research papers were accepted for the final programme at an overall acceptance rate of 16,7%. The accepted papers reflect research undertaken in Computer Science, Information Technology and Information Systems-related fields.

We would like to thank everyone who assisted with the programme. This includes the SAICSIT 2021 Programme Committee, members of the SAICSIT 2021 organising committee and staff from Unisa Press who handled the publication of the proceedings.

We trust that the conference programme will lead to interesting and stimulating scholarly discourse among participants and authors.

SAICSIT 2021 chairs
Bobby Tait, Jan Mentz and Hugo Lotriet

.

Executive Summary

SAICSIT has been hosting its annual conference from 1987, while over the years the shape and nature of the society has changed. This year the School of Computing at the University of South Africa hosted this virtual conference. The theme of the conference was “Reimagining the Interconnected World”.

Kiren Kosygin Padayachee, Ken J. Nixon and Stephen Phillip Levitt open the volume with ``Determining Human Hand Performance with the Oculus Quest in Virtual Reality Using Fitts’s Law'. Increasingly Virtual Reality is finding its way into our lives. The research that is reported in this paper outlines the investigation of user performance using hand tracking as a key matrix

Christine Asaju and Hima Vadapalli look at `Estimating Students' Learning Affects: An Approach Based on the Recognition of Facial Emotion Expressions'. The new normal in the education space is the increased use of virtual online classes. The authors investigate the use of deep learning to identify emotional face changes of students to understand the students learning experience.

Geoffrey Lydall and Stephen Phillip Levitt tackle``Quality Impact of Accommodating Customer Requirements Through Plug-Ins and Configuration Files'. The authors focus on the customisation of specific aspects of Enterprise Resource Planning systems in the context of logistics.

Jecton Tocho Anyango and Hussein Suleman address `Supporting Trainee Teachers of Computer Science with Game Authoring Tools'. The authors investigate a unique aspect of game-based learning. The researchers developed a prototype game and then evaluated the user experience of the game.

In the final research paper, authors Mphumelelo Dhlamini, Irwin Brown and Grant Osterwyk explore the organizational barriers to collaboration in Big Data Analytics.

Prof Shawren Singh | School of Computing
Department of Information Systems | University of South Africa

Determining Human Hand Performance with the Oculus Quest in Virtual Reality Using Fitts's Law

Kiren Kosygin Padayachee
https://orcid.org/0000-0002-9277-4294
School of Electrical and Information Engineering, University of the Witwatersrand, South Africa
kiren.padayachee@gmail.com

Ken J. Nixon
https://orcid.org/0000-0001-5391-8147
School of Electrical and Information Engineering, University of the Witwatersrand, South Africa
ken.nixon@wits.ac.za

Stephen Phillip Levitt
https://orcid.org/0000-0001-6054-6134
School of Electrical and Information Engineering, University of the Witwatersrand, South Africa
Stephen.levitt@wits.ac.za

Abstract

The medium of virtual reality has become much more accessible to the general public in recent years. The Oculus Quest virtual reality headset, released in 2019, is a device that is much more affordable to the average consumer while still providing a wide range of capabilities. In this paper, a study is discussed which focused on investigating novice human performance using the hand-tracking capabilities of the Oculus Quest. Fitts's multidirectional tapping task as given in ISO/TS 9241-411 in the form of a virtual button-pressing application was implemented. This human performance is measured in a single metric called throughput that combines speed and accuracy. It is measured in bits per second (bps). A throughput of 3.7 bps was determined for hand tracking with the Oculus Quest. This is comparable to the lower end of the mouse throughput range (3.7–4.9 bps) and similar to other studies on hand tracking in virtual reality (3.5–4.1 bps). Metrics for error rates, accuracy and movement times and movement trajectory analysis are also provided. Suggestions are made for improving design on virtual reality interfaces based on the study results which show the target configurations that provide optimal performance. The study results show that the hand-tracking feature of the Oculus Quest is capable of allowing the user to use their hand quite naturally and efficiently as a computer-interaction device.

CCS Concepts: human-centred computing, virtual reality, laboratory experiments
Keywords: virtual reality, hand tracking, human performance, human–computer interaction, oculus quest

Introduction

The importance of virtual reality (VR) systems and the benefits they bring to humanity are becoming increasingly evident in the modern world. VR technology is advancing swiftly. Advanced systems with VR head-mounted displays (HMDs) that were once only found in research initiatives such as the NASA Ames VIEWlab Project (Fisher 2016) can now be found in the homes of ordinary people. VR is used in a variety of applications such as video games, 360° videos and training.

The rise of VR also brings new ways to interact with computer systems. Common devices used to interact with a virtual environment (VE) are the hand-held controllers that come bundled with the VR headset. This has typically been the case for commercial headsets such as the Oculus Rift (Oculus 2021c) and HTC Vive (HTC 2021). Much effort has also gone into allowing the use of hands as human–computer interaction (HCI) devices. VR gloves and gloveless hand tracking have allowed the use of one's hands to interact with VEs through gestural commands, touching of virtual objects or as simple pointing devices (Chen, Wu, and Lin 2015). The Oculus Quest headset (Oculus 2021c), released in 2019, has the built-in capability of tracking the user's hands enabling them as HCI devices and is the first HMD to do so without requiring additional hardware.

To determine the reliability of interactions in real tasks, these devices and techniques need to be evaluated for their usefulness. Fitts's Law is a model that is used to predict human movement towards a target (Fitts 1954). It has been used to determine performance in terms of speed and accuracy on many commonly used devices such as computer mice and touchscreens and to design more user-friendly interfaces (MacKenzie 2018). The aims of this study are to determine the performance of the Oculus Quest's hand-tracking feature using Fitts's Law and to discuss its strengths and weaknesses. Its performance will also be compared to other HCI devices and setups.

Background

Immersion and Presence in Virtual Reality

VR is used for entertainment in video games and 360° videos, and for training on medical procedures, construction, CNC milling machines, mining, aircraft and space missions (Aslandere et al. 2014; Grantcharov et al. 2004; Larsen et al. 2009; Lin et al. 2002; Loftin and Kenney 1995; Sacks, Perlman, and Barak 2013; Van Wyk and De Villiers 2009). The quality of the training from these applications and the resulting task performance will be directly affected by the degree of immersion and presence provided by the virtual training environment (Vora et al. 2002; Youngblut and Huie 2003).

In the pursuit to define and quantify the experience of VR, the two concepts of immersion and presence are frequently brought up in VR literature. Immersion can be

defined as the extent to which a computer display is able to deliver an inclusive, extensive, surrounding and vivid illusion of reality to the senses of a user interacting with a VE (Slater and Wilbur 1997). It can be quantifiable by the quality of information it provides to a user's senses. Presence can be defined as the state of consciousness of the user, directly affected by their sense of being in the VE. Users who are highly present in a VE should be more engaged in the virtual world compared to the surrounding physical world and should consider the displays as places rather than just images that they see. Presence is more centred on how well a user's behaviour in the VE will match their behaviour in similar situations in real life. The ideal VE will provide a user with a high level of immersion and a strong sense of presence to coax the user into believing that they are part of the virtual world.

Slater et al. (1996) explored the relationship between immersion, presence and task performance in VR. Users in their study were required to learn a set of moves for the game of Tri-Dimensional Chess in VR or by viewing it on a television screen and replicate the end state of the board on the real chess board. Two environments were also used, a realistic garden scenario and a plain environment, with the chess game suspended in a void. User performance in replicating the chess board was much better in subjects using VR rather than the television screen. User performance was also better in subjects who were exposed to the more realistic garden scenario. The results showed that greater immersion produced better task performance, supporting the idea that immersion improved comprehension and memorisation of the 3D environment and movement sequence. Slater et al. (1996) argued that better immersion would improve performance in certain tasks because of the high quantity and quality of information available. Presence, however, is more concerned with how well the match is between the user's behaviour inside the VE and in similar circumstances in the real world. Slater et al. (1996) state that a feeling of being present brings about more natural reactions to a situation which may or may not have anything to do with the level of task performance by the user.

Hand Tracking in Virtual Reality

Although controllers are more commonly used to interact with VEs, they are ultimately more hardware to manage and require some time to get accustomed to in order to operate effectively (Masurovsky et al. 2020). Hand tracking is a solution that offers a more natural way for the user to interact with a VE. The user is instinctively inclined to interact with the VE as they would in real life. A VE that accommodates such interactions would improve the immersion and presence experienced by the user and could also contribute to better task performance.

Much research has been done on allowing the user to use their hands as HCI devices. Devices such as the Rutgers Master II-ND force feedback glove (Bouzit et al. 2002) and the Manus Prime II motion capture glove (Manus 2021) are high quality VR gloves which come with haptic feedback capabilities and which are mainly used for research

and industrial purposes. Devices that capture human movements for use in VEs that are more affordable to the general public include the P5 Glove (Davison 2007), the Leap Motion Controller (Shao 2016) and Microsoft's Kinect (Chen, Wu, and Lin 2015). The Microsoft Kinect and Leap Motion Controller enable the user to interact with computer programs without wearing any additional hardware such as gloves allowing for more natural interactions. When paired with VR headsets, these devices allow users to interact with VEs with their hands, increasing the immersion and presence of the user (Slater and Wilbur 1997).

The Oculus Quest headset takes this concept one step further. It uses four fisheye monochrome cameras in conjunction with neural networks to construct a virtual representation of the user's hands in the VE (Han et al. 2020). This is the first VR headset that supports this hand-tracking capability without requiring additional hardware. As the Oculus Quest offers similar capabilities of being able to use the human hand as a pointing device, it would be relevant to compare its performance in a VE against the previously mentioned devices as well as other established devices such as the computer mouse.

Determining Human Performance Using Fitts's Law

Measuring task performance quantitatively is a way to effectively evaluate the usefulness of a device. The ISO 9241-9 (ISO 2000) and ISO/TS 9241-411 (ISO 2012) specifications are focused on evaluating the task performance of HCI pointing devices such as mice and joysticks. These specifications present suggestions for optimising the design of pointing devices and provide standardised tests to evaluate them. These tests evaluate the pointing devices on user performance, comfort and effort.

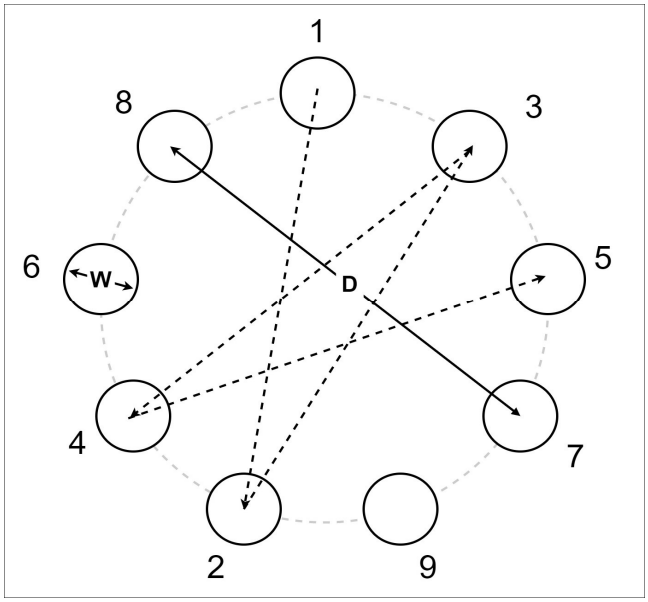
The 2D multidirectional tapping task (also referred to as Fitts's Test in this study) as given in ISO/TS 9241-411 makes use of Fitts's Law (Fitts 1954) to quantitatively evaluate user performance on the metric of throughput. Fitts's Law is a model used to predict human movement towards a target and has been widely used to study relationships for HCIs. It determines that the time taken to move towards a targeted area is related to the size of the target and the distance to it. The index of difficulty, ID, is a metric that quantifies how difficult a task is to complete and that is measured in bits. It is determined from the size of the target, W, and the distance, D, to move to the target (Equation 1). Movement time, MT, is linearly related to ID (Equation 2). The ratio of these values is the throughput, TP (Equation 3), measured in bits per second (bps).

$$ID = \log_2\left(\frac{D}{W} + 1\right) \quad (1)$$

$$MT = a + b \cdot ID \quad (2)$$

$$TP = \frac{ID}{MT} \quad (3)$$

The multidirectional tapping task has been widely used to evaluate pointing devices and tasks. Users are required to move from one target to the next in a predefined manner as depicted in Figure 1. The targets are spaced apart equally and are arranged in a circle (also referred to as Fitts's circle). By using multiple values of D and W in combination to achieve many values of ID, the movement time recorded from these movements can be used in linear regression methods to determine the overall throughput of a pointing device.



D: distance between targets
W: target width
Figure 1: The ISO/TS 9241-411 multidirectional tapping task with Fitts's circle

Measuring task performance is important when determining how effective an HCI device will be to use. Fitts's Law has been used on a variety of input devices to determine throughput which can then be compared to other devices. Computer mice, touchpads and joysticks have been evaluated using the ISO 9241-9 and ISO/TS 9241-411 tests (Douglas, Kirkpatrick, and MacKenzie 1999). Of particular interest is the measured throughput of the computer mouse, which has a throughput value ranging between 3.7 bps and 4.9 bps and which most other devices are compared to. The modern touchscreen yields a higher throughput at about 6 bps (Burno et al. 2015).

Fitts's Law in Virtual Environments

Fitts's Law has worked well for evaluating traditional HCI devices and has since been used in studies to evaluate devices used for interaction with VEs.

Teather et al. used the multidirectional tapping task to evaluate performance using passive haptic feedback in a Cave Automatic Virtual Environment (CAVE) in which the VE is projected onto the walls of a room-sized cube (Teather, Natapov, and Jenkin 2010). Fitts's circle was displayed in the VE with a transparent plastic panel positioned in front of the user to provide passive haptic feedback. The subjects each used a stylus to click on the targets in Fitts's circle. The throughput determined was 2.56 bps.

Pino et al. (2013) used the Kinect to allow their subjects to move a cursor on a monitor screen with their hands. A click was recognised by the users saying a sound which was then identified by the Kinect's microphones. The multidirectional tapping task was used for the evaluation. The 2D experiment yielded a throughput of 2.1 bps and the 3D experiment yielded 1.06 bps.

Joyce and Robinson conducted a similar experiment with users viewing Fitts's circle in a VE with the difference of clicking on its targets with their fingers instead (Joyce and Robinson 2017). This was enabled by using the Oculus Rift Development Kit 2 in conjunction with the Leap Motion Controller for hand tracking. The subjects touched targets in two separate conditions, one without any haptic feedback and one with a flat panel used to provide passive haptic feedback. The no-haptics condition yielded a throughput of 4.1 bps and the condition with passive haptic feedback yielded 4.7 bps. In addition, they found that the training time decreased with the haptic feedback and that the users preferred the passive haptics to interact with, as they felt it helped them perform faster and more accurately and that it also increased their sense of presence.

Mutasim et al. used the HTC Vive Pro Eye headset with a Leap Motion Controller for hand tracking (Mutasim, Batmaz, and Stuerzlinger 2021). They also used the multidirectional tapping task to determine throughputs for three techniques of clicking, pinching and dwelling on targets. The highest throughput was from the clicking technique which yielded around 4 bps.

Methodology

Equipment

Prior studies on VEs used numerous hardware such as the projectors in the CAVE. The experiments evaluating subjects using their hands used VR headsets in conjunction with the Leap Motion Controller. The first version of the Oculus Quest was chosen for this experiment as it has minimal hardware requirements in that only the headset is needed. The Oculus Quest has hand tracking built-in. This first version has a display refresh rate of 72 Hz giving the user a smooth viewing experience which increases immersion and presence. The Quest has a hand-tracking rate of 30 Hz (Han et al. 2020). The hand-tracking rate is noticeably lower than that of the Leap Motion Controller which typically operates at 120 Hz for hand tracking (Ultraleap 2021). Nevertheless, the hand-tracking feature of the Oculus Quest is still quite smooth and accurate in practise and therefore

acceptable for use in this study. Adequate lighting is required for the Oculus Quest to track hand movements accurately.

The Oculus Quest's high-accuracy controllers are used for the initial calibration of the experiment so that users can interact with the test application interface comfortably according to their height requirements.

Interface

A VR Android application for the Oculus Quest was developed using the Unity 3D engine (Unity Technologies 2021c) named VR Fitts's Test. This application is a heavily modified version of the HandsInteractionTrainScene from the sample framework provided by Oculus (2021b). The VE of the application consists of a ground, skybox and a central blue control panel as seen in Figure 2. The control panel is angled at 45° so that users do not have to bend their necks excessively (as would be the case if it were flat at 0°) and arm fatigue is reduced from holding their arms in mid-air (as would be the case if it were facing the user directly at 90°). The user is able to view virtual representations of the Oculus Touch controllers and blue virtual representations of their hands. The control panel has virtual red arcade buttons for the user to push with their virtual index fingers as input and an attached textbox for instructions and help. Pushing a virtual button all the way triggers a click sound to give the user audible feedback.

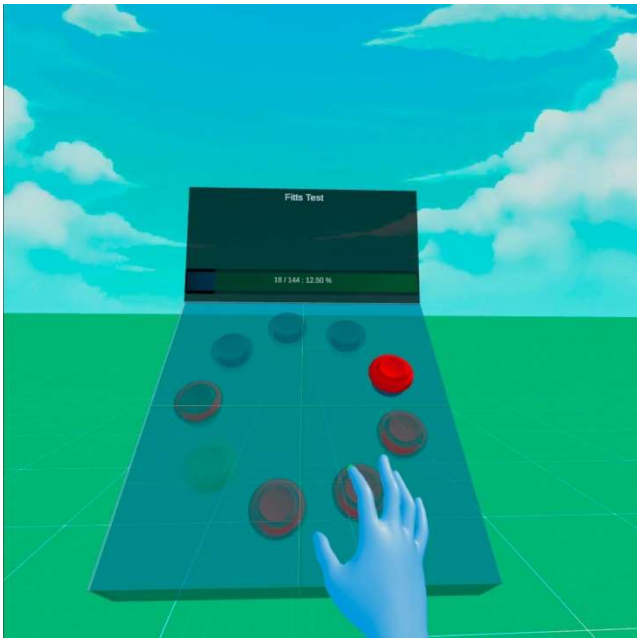


Figure 2: The VR Fitts's Test application user interface

The interface moves through various "scenes" for height calibration, inputs for the dominant hand, gender and age choices, and the multidirectional tapping task (named

as Fitts's Test within the application). These scenes all have different button layouts and help text. The user is given the choices of "Right" and "Left" for the dominant hand choice scene. For the gender choice scene, the options of "Male", "Female" and "Choose not to say" are given. Buttons labelled from zero to nine are shown for the age input scene, with a "Reset" button to erase what has been entered in case of a mistake and an "OK" button to confirm the entered age.

The user is given a practise round scene before moving on to the proper Fitts's Test scene and final completion scene. Both these scenes cycle through different configurations of the multidirectional tapping task.

Data Capturing

The VR Fitts's Test application captures the dominant hand, gender and age data entered by the user into memory during their interaction. It also captures all data into memory for the practise and proper Fitts's Test rounds. Each movement's start, end and target button centre 3D positions were captured and the movement time between start and end positions. The data for each circle configuration (see Figure 1) including distance between targets, target width and expected index of difficulty were also captured. Unique identifiers for the movement, trial, block and user are captured to determine the final throughput calculations. Finally, the index fingertip position, distance travelled, movement time and velocity for every frame in a movement's trajectory are captured. The hand-tracking rate on the Oculus Quest is lower than its display refresh rate so only unique trajectory positions are captured as they are updated.

Upon completion of the test, the above data is saved in an XML file on the Quest's onboard storage. The practise round data is excluded from this file. Supplementary files with a flat XML format for the movement, trajectory and circle configuration data are also saved for easier extraction of the data in the statistical analysis software.

Experimental Design

The multidirectional tapping task as given in ISO/TS 9241-411 is used in this application to determine user performance. There are a few variations of the way in which to design this test and also the way in which to collect, clean and analyse the data. Soukoreff and MacKenzie (2004) analysed many studies on Fitts's Law and provided guidance towards a standard for these tests. This study will take guidance from both of these sources. Sixteen configurations of the multidirectional tapping task circle as given in ISO/TS 9241-411 were determined from four distances and four button widths. When setting the circle diameter, it is important to remember that the target-to-target distance, D, as seen in Figure 1 will be less than this diameter. The button widths were determined by changing the Unity scale on an existing button model from the HandsInteractionTrainScene in the Oculus sample framework. The actual width of the button W was then determined using the Bounds functionality in Unity (Unity

Technologies 2021b). These four values of D and four values of W determine 16 values of ID (Equation 1).

Soukoreff and MacKenzie (2004) suggest an ID range of 2 bits to 8 bits. Lower ID values are simple to accommodate, but higher ID values require increasing D or reducing W (Equation 1). Making the circle too wide would result in forcing the user to move their whole body to reach from one end of the circle to the other. Reducing the button width to less than 2 cm had a negative impact on the user's experience in the pilot studies. The fingertip is much larger than the button at these sizes and the button is occluded by the user's finger. These are defined as the fat-finger problem and the occlusion problem respectively (Wigdor et al. 2007). Table 1 presents the chosen range of 1.701 bit to 4.808 bits which made the experience challenging while maintaining comfort.

Table 1: Circle configurations

Diameter (mm)	***D* (mm)**	**Button Scale**	***W* (mm)**	***ID* (bits)**
200	197	0.5	22	3.322
200	197	1.0	44	2.460
200	197	1.5	66	2.000
200	197	2.0	88	1.701
400	394	0.5	22	4.248
400	394	1.0	44	3.322
400	394	1.5	66	2.808
400	394	2.0	88	2.460
500	394	0.5	22	4.555
500	492	1.0	44	3.615
500	492	1.5	66	3.088
500	492	2.0	88	2.728
600	591	2.0	22	4.808
600	591	2.0	44	3.858
600	591	2.0	66	3.322
600	591	2.0	88	2.955

The subjects are required to go through all 16 configurations per block. These configurations are presented sequentially in a randomised order. One circle configuration is treated as a single trial for this study. The circle consists of n buttons for n movements per trial (n is configurable to allow for different practise and full-test designs). The subjects are required to click on all solid red highlighted buttons which appear in the order specified in Figure 1 and to avoid the transparent buttons and control panel surface in Figure 2. The first button click starts the timer for the user to complete the trial. A user needs to attempt to click all buttons in the sequence presented as fast and accurately as possible. If a button is missed and the user touches the control panel

surface or a transparent button instead, the panel will flash red and the next button in the sequence will be highlighted.

Keeping one's arm in mid-air and continuously moving it around for long periods is strenuous and induces fatigue. Therefore, after a trial is completed, a message appears next to the first button to indicate to the subject that they may take a short break if required. This aims to reduce fatigue in the study.

The user only uses their dominant hand for this test. The user starts the application viewing both virtual hands but then makes a choice for their dominant hand. The non-dominant hand is then removed from the user's view so that they can focus on the single virtual hand. The application is designed in such a way that collision detection with the buttons and control panel surface can only be triggered with the index finger of the chosen dominant hand. A small green sphere on the tip of the index finger is a visual cue to the user to only use that finger.

The application created for this test supports different circle configurations for which the distances, button widths, number of buttons and number of blocks can be set. This enabled different sequences for a practise round and the proper test.

The practise round consists of the user going through all 16 conditions for a single block. The circles consists of seven buttons from which seven movements will arise. This exposes the subject to a total of 112 movements. The practise round is important for subjects to become familiar with interacting with the VR environment, clicking on the buttons correctly, following the pattern of highlighted buttons and understanding the way in which a miss affects the flow of the test. This was important as all users were novices and some have never been exposed to VR at all before this study. Soukoreff and MacKenzie (2004) suggest that the subjects have enough training until they reach expert levels of performance before attempting the full test. Given the time constraints, this shorter practise round exposed the subjects to enough movements and circle configuration variation to help them get to an acceptable level of performance before starting the proper test. This will reduce learning effects which is not being considered for this study.

The full test employs a $3 \times 4 \times 4$ within-subjects design. The user is required to go through combinations of the four distances and the four button widths for a total of 16 conditions in a randomised sequence per block. The user will go through three of these blocks. The circles for the full test consist of nine buttons from which nine movements per trial will arise. This results in 432 total movements recorded per user.

Data Transformations

A movement's primary data consists of the movement time from start to finish, the start position, end position and the position of the target button centre. The adjustment for accuracy can be performed using this data to get the transformed values for effective

distance, De, and effective width, We, per subject per condition (Soukoreff and MacKenzie 2004). Looking at geometry of the movement in Figure 3, the distance for A, B and C are first calculated. These can then be used to calculate Dx and the effective distance travelled for the movement Ae (MacKenzie 2018). A negative value for Dx indicates an undershoot and a positive value indicates an overshoot.

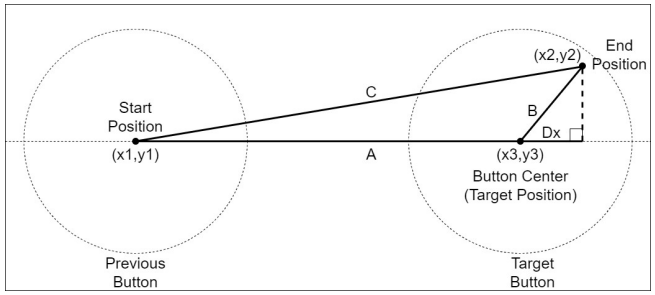


Figure 3: The geometry governing a movement from one target to the next

$$A = \sqrt{(x_3 - x_1)^2 + (y_3 - y_1)^2} \quad (4)$$

$$B = \sqrt{(x_3 - x_2)^2 + (y_3 - y_2)^2} \quad (5)$$

$$C = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (6)$$

$$Dx = \frac{C^2 - B^2 - A^2}{2A} \quad (7)$$

$$Ae = A + Dx \quad (8)$$

De is the mean of Ae for n movements for a single user for one condition across all blocks. Movement time, $\overline{MT}$, is also calculated as the average movement time for a single user for one condition across all blocks. SD_x is the standard deviation of Dx which is used to calculate We. De and We are then used to calculate the effective index of difficulty, *IDe*.

$$De = \frac{1}{n} \sum_{i=i}^{n} Aei \quad (9)$$

$$We = 4.133 \cdot SD_x \quad (10)$$

$$\overline{MT} = \frac{1}{n} \sum_{i=i}^{n} MT_i \quad (11)$$

$$IDe = \log_2 \left(\frac{De}{We} + 1\right) \quad (12)$$

The IDe and $\overline{MT}$ pairs are then used to determine the intercept and gradient for Equation 2 as well as the grand throughput TP (Equation 13) (Soukoreff and MacKenzie 2004). For y subjects and x conditions there are $y \times x$ pairs of IDe and $\overline{MT}$.

$$TP = \frac{1}{y}\sum_{i=1}^{y}\left(\frac{1}{x}\sum_{j=1}^{x}\frac{IDe_{ij}}{\overline{MT_{ij}}}\right) \quad (13)$$

Experimental Procedure

The experiment was conducted at the Chamber of Mines building located on the University of the Witwatersrand premises. A location in an open area in the building was chosen where it was well-lit by natural light. This requirement was necessary for the hand tracking to operate optimally on the Oculus Quest. The location had enough space for the subject to move around without bumping into anything. The Oculus Quest has a built-in "Guardian" system that allows setting up a virtual safety perimeter to prevent users bumping into any real world objects such as walls. The subjects could then be prepped to take part in the experiment.

An open invitation for participation was made to all students and staff at the School of Electrical and Information Engineering at the University of the Witwatersrand. The participants were verbally briefed on the procedures and aims of the experiment and also asked to read a participant information sheet with more details. If they wished to proceed with the experiment, a completed consent form was first required. The participant was then shown how to fit the Oculus Quest comfortably with a clear view of the VE. The controllers were then handed over and the participant was prompted to start the application called VR Fitts's Test.

For the initial calibration scene, the participants were prompted in-app to adjust the position of the control panel for their height and for a comfortable distance from their body to be able to interact with it. Once this step was completed, the participant passed the controllers back to the researcher. After that the participant would be able to see virtual representations of their hands and use their hands as interaction devices, being informed to click on the virtual arcade buttons with their index finger. They were then prompted in-app to enter their dominant hand, gender and age. The practise round then began. The first button of the trial informed the participants to take a break if they became tired. After the practise round, they proceeded to perform the full Fitts's Test. Both the practise round and the proper test had a progress bar indicating the completion percentage and how many movements had been completed out of the total.

A final scene indicated completion of the experiment and advised the participant to call the researcher to help them take off the headset. The researcher then checked the final scene for visual confirmation of when all files containing the participant's captured data had been successfully saved onto the device. The participant was then thanked for their participation. The participants did not receive any direct benefits and were all willing volunteers. The participants who completed the experiment were, however, offered further time to familiarise themselves with the device through preloaded games.

The average time spent in the application was 17 minutes. The fastest time was 12 minutes and the slowest was 31 minutes. The initial setup time including

explanations, signing of consent forms and fitting of the Oculus Quest typically took around 5–10 minutes. On average about 25–30 minutes was needed per participant for the full experiment procedure.

Results

Participants

A total of 23 participants took part in the study of which 13 were male and 10 were female. Three participants were left-handed. The average participant age was 29 and the ages ranged from 18 to 63. All the participants were able-bodied and had no impairments that would have worked against their favour. Some participants wore spectacles but this did not impair their performance as the Oculus Quest has enough space to accommodate users with glasses. Most participants had little to no experience with VR and can be considered novice subjects.

Data Reduction

A total of 9 936 movements were captured (23 subjects × 3 blocks × 16 trials × 9 movements per trial). Movements were flagged to indicate if the user missed the target. A movement was also flagged to indicate having low hand-tracking confidence at any point in its trajectory using the functionality from Oculus (Unity Technologies 2021a). Low confidence can arise from poor lighting, users obstructing the view of the cameras by blocking one hand with the other, holding the hand flat and perpendicular to the cameras so that no individual fingers can be distinguished, or moving too fast.

Hand-tracking errors did not have much effect on a participant's target-to-target movement data. These errors did, however, contribute to spikes in velocity and distance travelled in the trajectory data which were not possible by human means.

Target misses can sometimes have a significant effect on target-to-target movement data. Most misses will still have the user moving most of the distance towards a target and missing it by a small distance. This occurs fairly frequently for the smallest button widths. However, as mentioned by Soukoreff and MacKenzie (2004), subjects can sometimes double click on a target which registers a successful click for the current target but a miss for the next. Indeed it was observed that a number of subjects tried to click a button numerous times in its vicinity to make sure that the target button was clicked. This type of error creates a movement data capture with very low distance travelled and a very high effective button width which results in an extremely low *ID* value and a very short movement time. The movement following the double click is also very short as the new target is now adjacent to the user's current position. Leaving in these errors skew the results as there appears to be an extra distance close to zero contributing to the circle configurations.

Some participants also took much longer to complete the test. These participants took their time instead of rapidly moving which is required for the Fitts's Test. Some may

even have been taking a break in the middle of a trial, with one recorded movement taking over 13 seconds to complete. This is far off from the average movement time of 1 005 ms. When an error such as the double-click miss mentioned above occurred, almost impossibly low movement times are recorded with the lowest recording at 5 ms.

In summary:

- Too short movement times indicate double-click misses.
- Much longer times indicate the participant taking their time instead of clicking targets rapidly.
- Too short distances travelled indicate double-click errors.
- Too large distances from the button centre indicate that the participant's aim towards the target deviated too much.
- A double-click error can create a movement with an extremely small distance value and an extremely large button width value resulting in a very low value of ID.

Soukoreff and MacKenzie (2004) advise that trials in which final movement distance or time is more than three standard deviations away from the average be investigated. If these are deemed to be outliers then they should be removed. Based on the factors above, the recommendations from Soukoreff and MacKenzie have been extended to filter data on movement times, distance travelled and distance from the button centre on a per-condition basis for this study.

Since the IDe values are calculated on a per-user-per-condition basis, Joyce and Robinson (2017) only regarded a condition for a user as valid if at least a third of the total movements per condition are remaining after the initial data cleaning process. This step was also performed on this study's data set after cleaning for movement times, distances travelled and distances from the button centre. Starting off with 9 936 recorded movements and 368 potential pairs of IDe and $\overline{MT}$ for each condition across all users, a final total of 5 394 valid movements for 271 pairs of IDe and $\overline{MT}$ values remain for analysis.

Data Analysis

All calculations were done on a per-user-per-condition basis after cleaning the data. An increase in ID results in an increase in error rate, ER, and error in accuracy SD_x as seen in Table 2. The average movement time, $\overline{MT}$, is mostly affected by the distance component, D. Indices of difficulty are ordered from least to most difficult and labelled as condition i. Notice that there are some values of ID that are repeated. Even though these have the same value, they have different circle configurations and thus yield different results.

Table 2: Metrics per *ID* before data cleaning

i	ID (bits)	D (mm)	W (mm)	ER (%)	$\overline{MT}$ (ms)	SD_x (mm)
1	1.701	197	88	4.19	759	39
2	2.000	197	66	4.51	772	42
3	2.460	394	88	4.99	911	75
4	2.460	197	44	14.81	785	58
5	2.728	492	88	6.76	1068	98
6	2.808	394	66	8.21	949	93
7	2.955	591	88	8.05	1285	134
8	3.088	492	66	11.92	1059	125
9	3.322	591	66	12.88	1182	164
10	3.322	394	44	20.29	980	120
11	3.322	197	22	31.40	831	79
12	3.615	492	44	19.48	1076	152
13	3.858	591	44	20.77	1121	207
14	4.248	394	22	36.07	948	157
15	4.555	492	22	36.23	1092	204
16	4.808	591	22	38.32	1254	248

Figure 4 illustrates that as the ID increases, so does the error rate. Note, however, that it is not perfectly ordered even though there are increasing values of ID. The noticeable spike in error rate for $ID_{i=4}$ is mainly due to the smaller value of W at 44 mm. This spike in error is observed again for $ID_{i=11}$ for the smallest value of W at 22 mm. Indeed the biggest error rates are observed when the button width is the smallest at 22 mm for $ID_{i=11}$, $ID_{i=14}$, $ID_{i=15}$ and $ID_{i=16}$. Smaller button widths seem to have a larger effect on error rates than larger distances do.

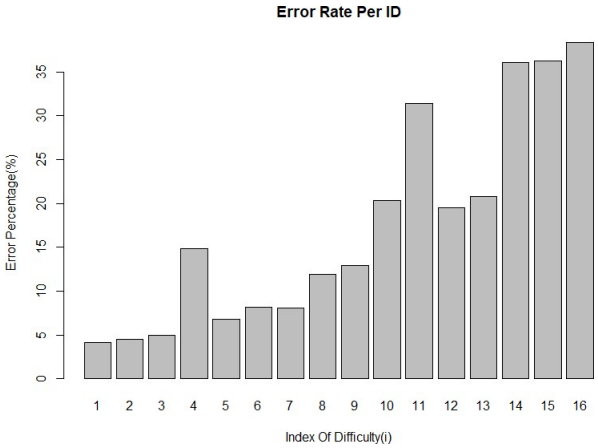


Figure 4: Error rate per *ID*

The error in accuracy SD_x is also observed to increase with increasing ID in Figure 5. A noticeable exception to this is for $ID_{i=11}$. This condition is comprised of the smallest distance and button width. Remember that for a larger button width, the subject can "cheat" by deliberately clicking anywhere between the edge of the button to the centre instead of aiming for the centre (Soukoreff and MacKenzie 2004). A smaller button width would in theory increase the accuracy as, even if the subject clicks on the edge of the button, the accuracy registered for We would be much better than clicking on the edge of a larger button. Larger distances and ID values certainly increase the error in accuracy even if the button width is kept small as seen in Figure 5. However, for the case of $ID_{i=11}$, the smaller distance to travel and small button width allow for easier targeting of the button and therefore better accuracy.

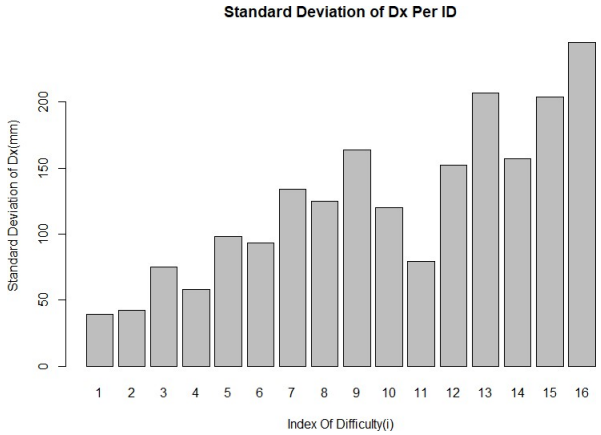


Figure 5: Standard deviation of *Dx* per *ID*

Movement time appears to be affected mainly by D. It can be observed in Figure 6 that $ID_{i=7}$, $ID_{i=9}$, $ID_{i=13}$ and $ID_{i=16}$ all have the largest D value at 591 mm contributing to the largest average movement times. $ID_{i=5}$, $ID_{i=8}$, $ID_{i=12}$ and $ID_{i=15}$ at the second largest value of D at 492 mm contribute to the next four largest average times and so forth.

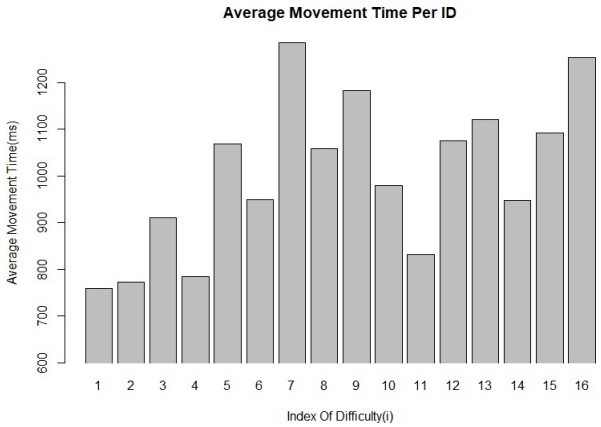


Figure 6: Average movement time per *ID*

Least-squares linear regression is used to determine if there is a linear relationship between MT and ID. It can be seen in Figure 7 that movement time increases with increasing ID after cleaning the data. Of note is the intercept at 447 ms. Soukoreff and MacKenzie (2004) mention as a rough guideline that the intercept should lie between −200 ms and 400 ms. This is, however, only for expert subjects. For this study, the participants can be regarded as novices as the practise round was limited and not enough to advance them to expert status before the proper test. It could also be attributable to the fact that much larger distances are used for this test requiring more time for movement (as compared to distances travelled for a mouse in pixels) or the lower 30 Hz sampling of the hand tracking on the Oculus Quest. In addition, no cleaning of the trajectory data is being performed. When users click a button, their hand goes straight through the button and immediately activates the next target as in Figure 8. There is some extra time taken to slow down and reverse direction to the next target after a successful click on the current button. This will slightly increase movement times. This extra movement after the click can be seen below the plane of the control panel surface in an example recorded trajectory in Figure 9 and could potentially be considered reaction time for removal from the trajectory data as done in the study by Soukoreff and MacKenzie (2004). Alternatively, it could also be remedied by using haptics (Joyce and Robinson 2017; Teather, Natapov, and Jenkin 2010). Homing and dwell time is not applicable to this study owing to the fact that the user is constantly moving from target to target and is not required to hover in the vicinity of the target (Soukoreff and MacKenzie 2004).

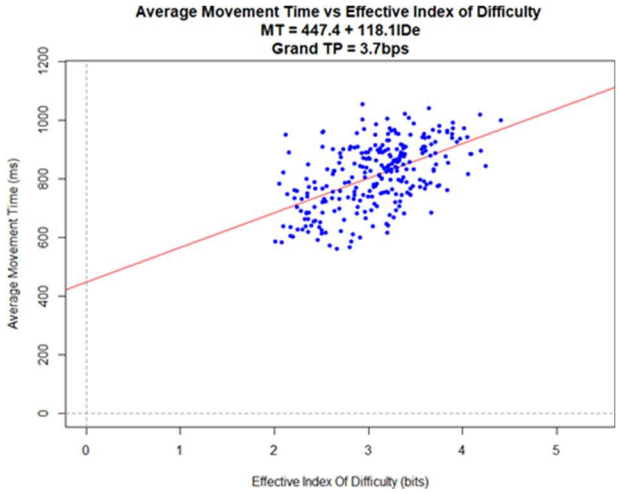


Figure 7: Average movement time vs effective index of difficulty, *IDe*

Using the average of the effective index of difficulty values $\overline{IDe}$ = 3.081 bits, the ( )
standard error of the intercept is $s.e.(a)$ = 311.68 ms. The F-statistic shows that the regression model MT = 447.4 + 118.1 IDe is a statistically significant fit for the data ($F_{1,271}$ = 108.2, $p < 2.2 \times 10^{-16}$).

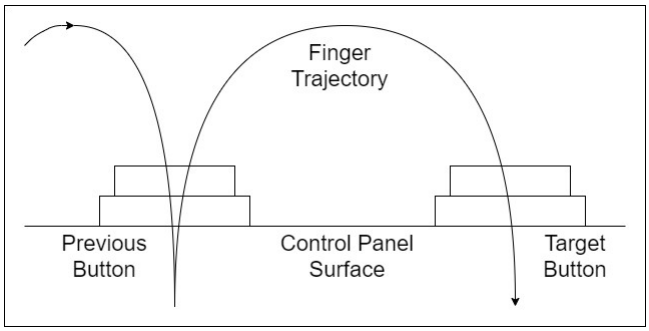


Figure 8: Finger trajectory over time

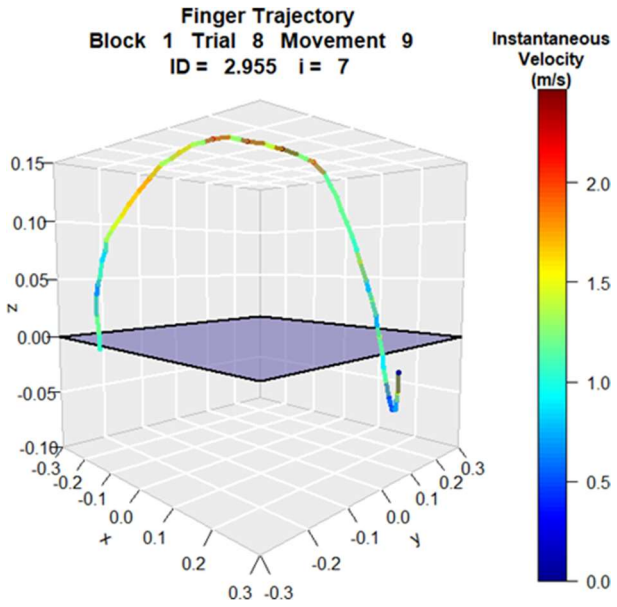


Figure 9: A typical finger trajectory for a single movement relative to the control panel plane

Because the subject's finger is still moving after clicking a target, the velocity profile of a movement trajectory will never be stationary at the beginning. It will always start at a non-zero velocity, reduce as the user changes direction, and then accelerate towards the next target, slowing down again to click the target accurately. Because the subject's finger goes through the button, the end velocity of the movement is again low but non-zero. This can be seen in Figure 10 and is similar to the velocity profiles plotted by Joyce and Robinson (2017). This was plotted from velocity data across all movements for a single trial for a subject.

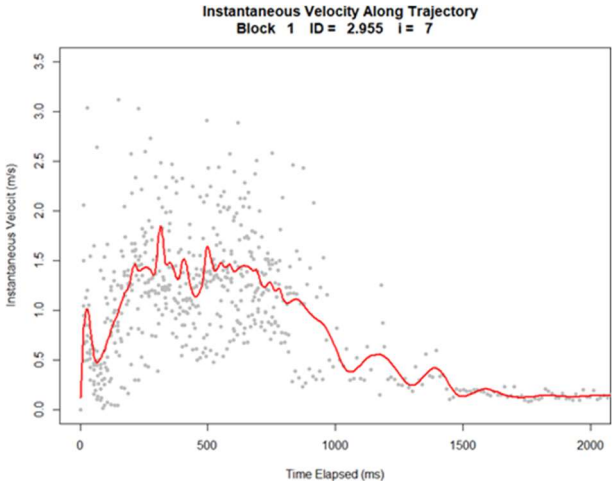


Figure 10: The typical velocity profile for movement trajectories

Discussion

Fitts's Law is important to consider when designing user interfaces, whether it be physical control panels and devices, operating systems, websites or VR interfaces. Designing interfaces to have low *ID* values between components will make working on it easier, more useful and more user-friendly. Studies such as this one therefore offer a wealth of information to consider when designing these interfaces.

Ultimately, the goal of a Fitts's Test is to determine the throughput of a device. For hand tracking with the Oculus Quest, this study determined a throughput of 3.7 bps (Figure 7). This is lower than the average throughput for the mouse (3.7–4.9 bps) but higher than the isometric joystick (1.6–2.55 bps) and the touchpad (0.99–2.9 bps) (Soukoreff and MacKenzie 2004).

The 2D and 3D experiments with the Microsoft Kinect by Pino et al. (2013) yielded a throughput of 2.1 bps and 1.06 bps, respectively. Teather et al. (2010) determined 2.37 bps without haptic feedback and 2.56 bps with haptic feedback for a stylus in a CAVE setup. Joyce and Robinson (2017) determined 4.1 bps without haptic feedback and 4.7 bps with haptic feedback using the Leap Motion Controller for hand tracking with the Oculus Rift Development Kit 2. Mutasim, Batmaz and Stuerzlinger (2021) determined a throughput of around 4 bps using the Vive controller with the HTC Vive Pro Eye for clicking on targets and around 3.5 bps with the Leap Motion Controller enabling a pinching action to select targets. MacKenzie (2018) determined 6.85 bps for a touchscreen. Burno et al. (2015) determined around 6 bps for a touchscreen and around 2 bps for hand tracking using the Leap Motion Controller. Touchscreens seem to yield the best throughput because of direct input from the user instead of indirect input such as through the mouse pointer and mouse click (MacKenzie 2018).

The throughput determined from this study is comparable to the lower end for the mouse, the no-haptics condition in Joyce and Robinson's study and the throughput values for clicking and pinching in the study by Mutasim et al. It is much higher than the throughput values yielded from the studies by Pino et al. A throughput of 3.7 bps therefore proves that the Oculus Quest's hand-tracking capabilities are effective. This, plus the fact that the Oculus Quest requires no extra hardware for hand-tracking, makes it an even more appealing device. It might be worthwhile repeating this study using haptic feedback to test if this also boosts throughput, as seen in other studies. Of note for this study and the throughput determined are the following:

- Physical distances between targets are much larger than most other studies in the range of 200–600 mm, which are much larger than the distances needed for moving a mouse on a screen (pixels versus millimetres) contributing to larger movement times and therefore reduced throughput (Equation 3).
- The participants were not given the chance to reach expert levels before the test and were tested as novices, given just bare minimum practise to get used to the interface and test proceedings. It might be worth conducting the test with expert level users to test if the level of expertise increases throughput.
- Most Fitts's tests require users to travel in a straight line to targets. This test required the participants to move more naturally resulting in arc movements. The 2D point-to-point distance was used rather than the full 3D trajectory distance. Using the 3D trajectory distance might yield a larger *ID* range and might contribute to a better throughput (Equation 1). Using the full 3D trajectory distance might, however, require some modifications of Fitts's Law to accommodate 3D movements.
- The hand-tracking feature of the Oculus Quest is relatively new, having first been added to the platform towards the end of 2019. Improvements in the software and neural network used to provide this feature would surely improve throughput over time.

It must be noted that using hand tracking on the Oculus Quest does not come without its shortcomings. There are occasional hand-tracking issues, the most common owing to poor lighting. This could be improved by using better cameras or image processing. There is the issue of hands occluding each other. Perhaps more training of the neural networks is required to remedy this. There is also the issue of fatigue. A user interacting with a VR interface in mid-air is going to inevitably experience fatigue. Indeed this was noticed during the test and there is a questionnaire provided by ISO/TS 9241-411 that evaluates a user's fatigue. This could be used in further studies to gain more insight on this issue. A haptic feedback solution could possibly ease the fatigue experienced by users. It could also be a potential solution for VR interface designers to incorporate breaks (such as in this test) or workflows that reduce fatigue.

It was noted that the ID range for this study was limited to 1.701–4.808 bits even though the recommendation is for a range of 2–8 bits (Soukoreff and MacKenzie 2004). This recommended range is easier to implement when the pointing device is more precise. A mouse pointer tip is a single pixel so much smaller targets are feasible, as compared to a fingertip which is in the order of millimetres. In the case of VR with hand tracking, as in this study, button widths smaller than 2 cm could increase the ID range but could have a severe impact on user experience because of the fat-finger problem and the occlusion problem (Wigdor 2007).

It can be observed from Table 2 and Figure 4 that the circle configurations with the smallest button width yield the most errors, which is an indication that these scales are not suitable for the context of hand tracking in VR. It might be possible to increase this ID range in future studies by changing the virtual representation of the hand to something more precise such as a skeleton hand made out of fine lines. It might also be worth conducting the same test with one of the Oculus Touch controllers for comparison since this is effectively used as a mouse for this platform.

Conclusions

This study investigated user performance using the hand-tracking feature on the Oculus Quest. Fitts's multidirectional tapping task as given in ISO/TS 9241-411 was used to evaluate performance as throughput which combines speed and accuracy. The final throughput determined was at the same level as the lower end of mouse throughput and similar to using hand tracking in VR with no haptic feedback. It was higher than other devices such as isometric joysticks, styli, touchpads and the Microsoft Kinect, but not as high as throughputs for touchscreens.

Of note is that this throughput was determined with novice subjects and it is expected to increase with expert subjects. The throughput determined confirms that the hand-tracking capabilities of the Oculus Quest allow the user to use their hands naturally, efficiently and accurately as pointing devices. The results from this study can be used to design better VR interfaces, especially when hand tracking can be used. It is noted from movement times, error rates and accuracy data that for this particular situation of hand tracking in VR maximum efficiency is achieved when the size of interface components are larger than the user's finger and spaced more closely together. Even though the throughput determined is lower than some devices such as the mouse and touchscreen, it is worth noting that these devices are used most advantageously for appropriate situations.

Evaluating devices using Fitts's Law yields great insights into developing better HCI devices and techniques such as hand tracking, which ultimately benefits the end user.

References

Aslandere, Turgay, Danile Dreyer, Frieder Pankratz, and René Schubotz. 2014. “A Generic Virtual Reality Flight Simulator.” In *Virtuelle und Erweiterte Realität, 11. Workshop der GI-Fachgruppe Tagung Band,* 1–13. Shaker.

Bouzit, Mourad, George Popescu, Grigore Burdea, and Rares Boian. 2002. “The Rutgers Master II-ND force feedback glove.” In *Proceedings 10th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. HAPTICS 2002*, 145–52. IEEE.

Burno, Rachael A., Bing Wu, Rina Doherty, Hannah Colett, and Rania Elnaggar. 2015. “Applying Fitts’ Law to Gesture Based Computer Interactions.” *Procedia Manufacturing* 3: 4342–49. https://doi.org/10.1016/j.promfg.2015.07.429.

Chen, Wei-Lun, Chih-Hung Wu, and Chang Hong Lin. 2015. “Depth-Based Hand Gesture Recognition Using Hand Movements and Defects.” In *2015 International Symposium on Next-Generation Electronics (ISNE)*. IEEE, 1–4. https://doi.org/10.1109/ISNE.2015.7132005.

Davison, Andrew. 2007. “The P5 Glove.” *Pro Java™ 6 3D Game Development: Java 3D™, JOGL, JInput, and JOAL APIs* 349–73.

Douglas, Sarah A., Arthur E. Kirkpatrick, and I. Scott MacKenzie. 1999. “Testing Pointing Device Performance and User Assessment with the ISO 9241, Part 9 Standard.” In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 215–22. https://doi.org/10.1145/302979.303042.

Fisher, Scott S. 2016. “The NASA Ames VIEWlab Project – A Brief History. *Presence: Teleoperators and Virtual Environments* 24 (4): 339–48. https://doi.org/10.1162/PRES_a_00277.

Fitts, Paul M. 1954. “The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement.” *Journal of Experimental Psychology* 47 (6): 381. https://doi.org/10.1037/h0055392.

Grantcharov, Teodor P., Viggo B. Kristiansen, Jørgen Bendix, Linda Bardram, Jacob Rosenberg, and Peter Funch-Jensen. 2004. “Randomized Clinical Trial of Virtual Reality Simulation for Laparoscopic Skills Training.” *British Journal of Surgery* 91 (2): 146–50. https://doi.org/10.1002/bjs.4407.

Han, Shangchen, Beibei Liu, Randi Cabezas, Christopher D. Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. 2020. “MEgATrack: Monochrome Egocentric Articulated Hand-Tracking for Virtual Reality.” *ACM Transactions on Graphics* 39 (4): 87–1. https://doi.org/10.1145/3386569.3392452.

HTC. 2021. “VIVE United States/Discover Virtual Reality beyond Imagination.” Accessed 25 April 2021. https://www.vive.com/us/.

ISO (International Organization for Standardization). 2000. *ISO 9241-9:2000, Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) – Part 9 : Requirements for Non-Keyboard Input Devices*. Geneva: ISO.

ISO (International Organization for Standardization). 2012. *ISO/TS 9241-411:2012, Ergonomics of Human-System Interaction – Part 411: Evaluation Methods for the Design of Physical Input Devices*. Geneva: ISO.

Joyce, Richard D., and Stephen Robinson. 2017. “Passive Haptics to Enhance Virtual Reality Simulations.” In *AIAA Modeling and Simulation Technologies Conference*. 1313. https://doi.org/10.2514/6.2017-1313.

Larsen, Christian R., Jette L. Soerensen, Teodor P. Grantcharov, Torur Dalsgaard, Lars Schouenborg, Christian Ottosen, Torben V. Schroeder, and Bent S. Ottesen. 2009. "Effect of Virtual Reality Training on Laparoscopic Surgery: Randomised Controlled Trial." *BMJ* 338. https://doi.org/10.1136/bmj.b1802.

Lin, Fuhua, Lan Ye, Vincent G. Duffy, and Chuan-Jun Su. 2002. "Developing Virtual Environments for Industrial Training." *Information Sciences* 140 (1–2): 153–70. https://doi.org/10.1016/S0020-0255(01)00185-2.

Loftin, R. Bowen, and P. Kenney. 1995. "Training the Hubble Space Telescope Flight Team." *IEEE Computer Graphics and Applications* 15 (5): 31–37. https://doi.org/10.1109/38.403825.

MacKenzie, I. Scott. 2018. "Fitts' Law." *Handbook of Human–Computer Interaction* 1: 349–70. https://doi.org/10.1002/9781118976005.ch17.

Manus. 2021. "Manus/Prime II for Mocap." Accessed 25 April 2021. https://www.manus-vr.com/mocap-gloves.

Masurovsky, Alexander, Paul Chojecki, Detlef Runde, Mustafa Lafci, David Przewozny, and Michael Gaebler. 2020. "Controller-Free Hand Tracking for Grab-and-Place Tasks in Immersive Virtual Reality: Design Elements and their Empirical Study." *Multimodal Technologies and Interaction* 4 (4): 91. https://doi.org/10.3390/mti4040091.

Mutasim, Aunnoy K., Anil Ufuk Batmaz, and Wolfgang Stuerzlinger. 2021. "Pinch, Click, or Dwell: Comparing Different Selection Techniques for Eye-Gaze-Based Pointing in Virtual Reality." https://doi.org/10.1145/3448018.3457998.

Oculus. 2021a. "HandsInteractionTrainScene Sample Scene/Oculus Developers." Accessed 25 April 2021. https://developer.oculus.com/documentation/unity/ unity-sf-handtracking/.

Oculus. 2021b. "Oculus Quest: All-in-One VR Headset/Oculus." Accessed 25 April 2021. https://www.oculus.com/quest/.

Oculus. 2021c. "Oculus Rift S: VR Headset for VR Ready PC's/Oculus." Accessed 25 April 2021. https://www.oculus.com/rift-s/.

Pino, Alexandros, Evangelos Tzemis, Nikolaos Ioannou, and Georgios Kouroupetroglou. 2013. "Using Kinect for 2D and 3D Pointing Tasks: Performance Evaluation." In *International Conference on Human–Computer Interaction*, 358–367. https://doi.org/10.1007/978-3-642-39330-3_38.

Sacks, Rafael, Amotz Perlman, and Ronen Barak. 2013. "Construction Safety Training Using Immersive Virtual Reality." *Construction Management and Economics* 31 (9): 1005–17. https://doi.org/10.1080/01446193.2013.828844.

Shao, Lin. 2016. "Hand Movement and Gesture Recognition Using Leap Motion Controller." *Virtual Reality, Course Report*.

Slater, Mel, Vasilis Linakis, Martin Usoh, and Rob Kooper. 1996. "Immersion, Presence and Performance in Virtual Environments: An Experiment with Tri-Dimensional Chess." In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, 163–172. https://doi.org/10.1145/3304181.3304216.

Slater, Mel, and Sylvia Wilbur. 1997. "A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments." *Presence: Teleoperators and Virtual Environments* 6 (6): 603–16. https://doi.org/10.1162/pres.1997.6.6.603.

Soukoreff, R. William, and I. Scott MacKenzie. 2004. “Towards a Standard for Pointing Device Evaluation, Perspectives on 27 Years of Fitts’ Law Research in HCI.” *International Journal of Human–Computer Studies* 61 (6): 751–89. https://doi.org/10.1016/j.ijhcs.2004.09.001.

Teather, Robert J., Daniel Natapov, and Michael Jenkin. 2010. “Evaluating Haptic Feedback in Virtual Environments Using ISO 9241–9.” In *2010 IEEE Virtual Reality Conference (VR)*, 307–308. IEEE. https://doi.org/10.1109/VR.2010.5444753.

Ultraleap. 2021. “Leap Motion Datasheet A.” Accessed 25 April 2021. https://www.ultraleap.com/datasheets/Leap_Motion_Controller_Datasheet_April_2020.pdf.

Unity Technologies. 2021a. “Hand Tracking in Unity/Oculus Developers.” Accessed 25 April 2021. https://developer.oculus.com/documentation/unity/unity-handtracking/.

Unity Technologies. 2021b. “Unity – Scripting API: Bounds.” Accessed 25 April 2021. https://docs.unity3d.com/ScriptReference/Bounds.html.

Unity Technologies. 2021c. “Unity Real-Time Development Platform/3D, 2D VR and AR Engine.” Accessed 25 April 2021. https://unity.com/.

Van Wyk, Etienne, and Ruth de Villiers. 2009. “Virtual Reality Training Applications for the Mining Industry.” In *Proceedings of the 6th International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa,* 53–63. https://doi.org/10.1145/1503454.1503465.

Vora, Jeenal, Santosh Nair, Anand K. Gramopadhye, Andrew T. Duchowski, Brian J. Melloy, and Barbara Kanki. 2002. “Using Virtual Reality Technology for Aircraft Visual Inspection Training: Presence and Comparison Studies.” *Applied Ergonomics* 33 (6): 559–70. https://doi.org/10.1016/S0003-6870(02)00039-X.

Wigdor, Daniel, Clifton Forlines, Patrick Baudisch, John Barnwell, and Chia Shen. 2007. “Lucid Touch: A See-Through Mobile Device.” In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology,* 269–78. https://doi.org/10.1145/1294211.1294259.

Youngblut, Christine, and Odette Huie. 2003. “The Relationship between Presence and Performance in Virtual Environments: Results of a VERTS Study.” In *IEEE Virtual Reality, 2003. Proceedings*, 277–278. IEEE.

Estimating Students' Learning Affects: An Approach Based on the Recognition of Facial Emotion Expressions

Christine Asaju
https://orcid.org/0000-0003-2728-6806
Computer Science Department,
University of the Witwatersrand, South Africa
Chrisamaju02@gmail.com

Hima Vadapalli
https://orcid.org/0000-0001-9040-3601
Computer Science Department,
University of the Witwatersrand, South Africa
Hima.vadapalli@wits.ac.za

Abstract

Online education has experienced rapid development owing to its significance as a potential solution to teaching and learning under the critical conditions caused by Covid-19. A major obstacle in this form of learning is that online classes lack direct, timely and effective communication and feedback to teachers. The use of machine learning algorithms for estimating facial emotion expressions of students during teaching sessions has garnered interest from researchers in the last decade; however, there has been no or limited feedback mechanisms incorporated into these models. In the present study, we explored the use of deep learning to identify emotional changes in students' faces and use them to estimate the learning affect experienced by the students. This involved implementing a CNN-BiLSTM model for emotion expression recognition and mapping of identified emotions into positive, negative and neutral learning affects, and for further affect analysis. This model was trained, validated and tested by using the extended DISFA database. A test accuracy of 92 per cent on a sample size of 2 274 was reported. The classified emotions were then mapped into learning affects, based on mappings provided in the literature. The model was further tested on live samples (collected in a laboratory set-up) to ascertain the validity of the mappings. It is envisaged that the analysis of the learning affects through facial emotion changes will potentially pave the way for timely and appropriate feedback to teachers on the learning affect experienced by students, potentially improving the feedback mechanism of the existing e-learning platforms.

CCS concepts: computing methodologies, machine learning, machine learning approaches, neural networks

Keywords: online learning, facial emotion recognition, learning affects analysis, deep learning

Introduction

Most learning institutions globally have not been able to be fully operational owing to the Covid-19 pandemic and have adopted online teaching and learning as one of their main tools. During these online teaching sessions, it is more so important for teachers to get an insight into their students' learning process. Some of the methods adopted for affect analysis include quizzes, open-ended questions, drag-and-drop activities, peer evaluations and reviews, surveys, and gamification.

Although these methods have been widely used, they are mostly offline and lack the ability to provide real-time feedback to teachers during a teaching and learning session. Real-time tracking and analysis of facial emotion expressions exhibited by students have the potential to fill this gap. It was opined by Phillips (1993) that facial emotion expressions are the most appropriate source of non-verbal communication, which accounts for up to 93 per cent of the impact of any verbal message. This has triggered the use of facial emotion expression recognition as an important area of research for e-learning set-ups. Previous studies have shown that the emotional states and motivation of students influence the learning process either directly or indirectly (Pekrun 1992). Diverse studies further proved that facial expression recognition can be used to evaluate the emotional states of students in an online learning environment and that the various learning affects can be inferred from the facial emotions expressed (Spector et al. 2014; Tyng et al. 2017).

Facial emotion expression recognition is an important area in machine learning. Facial emotion expression recognition systems enable the application of emotion-related knowledge to improve human and computer communication and to have a more satisfying user experience. A good grasp of human emotions by the computer can give one access to several opportunities and many applications. Some of the applications include targeted advertising campaigns, ATM payments, e-learning, healthcare, and online gaming. Devices that sense human activities can provide an enriched interaction and can also enrich the users' experience. Subsequently, facial emotion expression has been proved to be universal (Power and Dalgleish 1999). Power and Dalgleish (1999) opined that humans have a range of emotions, identifying seven major categories, namely, anger, disgust, fear, happiness, sadness, surprise, and neutral. For a teacher to teach effectively to meet the students' needs during a class, they must be able to quickly access the state of their audience (Ferdig et al. 2020; Klein and Celik 2017). Getting a good understanding of students' affects during learning continues to be an important objective to teachers. Students experience different emotions either cognitively or affectively when they are learning. The specific problem this work seeks to solve is estimating the students' learning affects during online class with the use of facial emotion expression recognition. The estimated learning affects can then provide insights into student comprehension that can act as feedback to teachers, which is a potential improvement to the existing platforms.

Given the success of deep learning, it is imperative that more research be carried out on the usage of deep learning to improve the interaction between humans and computers, especially in online-learning scenarios. The common deep-learning techniques used for facial emotion recognition are the convolutional neural network (CNN) and long short-term memory (LSTM). However, the traditional deep-learning approach to feature extraction has its shortcomings. One such limitation is that a model is trained using particular data for a specific task. For new tasks, the model has to be rebuilt. The adoption of transfer learning can improve the performance of a deep-learning model by transforming known knowledge learned from other related data (Dhankhar 2019; Khanzada, Bai, and Celepcikay 2020; Wahab, Khan, and Lee 2019). Several pretrained CNN architectures have been proposed by researchers. Some of these pretrained models include: ResNet50, 101, 152, VGG16, 19, MobilenetV2, InceptionResnetv2, and Densenet, which are trained on the ImageNet data set (Marcelino 2018; Siami-Namini , Tavakoli, and Namin 2019).

In order to achieve the above objectives, this research proposes the use of transfer learning for extracting discriminate features of the facial data and a bidirectional LSTM for capturing and encoding both spatial and temporal features of the facial emotion expression data. The study therefore fine-tunes the ResNet50 pretrained network for extracting features of facial image samples obtained from the extended Denver Intensity of Spontaneous Facial Action (DISFA+) database. The extracted features of facial images are used for classifying the various facial emotions using an enhanced recurrent neural network (a bidirectional LSTM) and for mapping the classified emotions to various learning affects for further analysis.

Related Works

Over the last decade, attempts were made to recognise and analyse facial expressions for use in both online and face-to-face learning environments using diverse approaches. Ayvaz, Gürüler and Devrim (2017) proposed an information system that detects the emotional states of learners; this is the information about the instant and weighted emotional states that were based on their facial expressions. Their work intended to create a formal interactive virtual environment. Experiments included the use of the k-nearest neighbour and the support-vector machine (SVM), which produced a best accuracy of 96.38 per cent and 98.24 per cent respectively.

Sun et al. (2017) employed an emotion detection module applicable to e-learning. They used CNNs to detect emotion in an e-learning setting and their model achieved an accuracy of 84.55 per cent. Mukhopadhyay et al. (2020) assessed the emotions of learners and identified the emotional changes that occur during learning. Basic emotions were classified using a CNN model which was further used to estimate a learners' state of the mind. An accuracy of 65 per cent and 62 per cent for the emotion classification and states of mind identification was reported.

Ray and Chakrabarti (2012) focused on the detection of any changes in emotion during the learning process using biophysical signals. A neural network was used to classify learners' emotions and to predict their learning style. The work proposed to indicate the effectiveness of facial emotion recognition for identifying the affective level of a student using the SVM and presented an accuracy regarding the average gain for their sample as 1.3693. The average gain result is further applied for identifying course delivery patterns according to the students' learning styles. Dewan, Murshed and Lin (2019) explored the use of facial images in estimating the engagement levels of students in an online environment. The use of InceptionNet achieved accuracies of 36.5 per cent, 47.1 per cent, 70.3 per cent, and 78.3 per cent respectively for the engagement levels for boredom, engagement, confusion, and frustration.

Megahed and Mohammed (2020) considered the integration of a CNN and a fuzzy system for estimating facial expressions and predicting the next learning level respectively. A fuzzy system was set up to take in the extracted facial expression states estimated by the CNN. Pise, Vadapalli and Sanders (2020) proposed the use of a temporal relational network for identifying students' facial emotion changes during an e-learning session to determine the students' learning affects and levels of engagement. The study considered the use of a single scaled and multiscale temporal relation network compared with a multilayer perceptron as a baseline model. The DISFA data set was used for the experiments and achieved an accuracy of 92.7per cent, 89.4 per cent and 86.6 per cent on the multi-scale TRN, single scale TRN and MLP respectively.

Dewan, Murshed and Lin (2019) proposed a facial emotion recognition method in which a bank of a discrete Hidden Markov Model is learned using the Baum–Welch algorithm on selected coefficients of singular value decomposition. Six academic emotions such as delight, disappointment, boredom, confusion, engagement, and frustration were recognised. The CMU-PIE data set was used for evaluating the model and tested both class-dependent and class-independent cases reporting an accuracy of 68 per cent and 90 per cent respectively. De Carolis et al. (2020) presented the development of a facial emotion expression recognition system that can recognise cognitive emotions in a distance education domain. The system developed classified a few cognitive emotions such as enthusiasm, interest, surprise, boredom, perplexity, frustration, and neutral. Combinations of features and best algorithms for classification problems as action units (AUs) and gaze features were reported after extensive experimentation. A multi-class SVM was reported to be producing the best accuracy on the data set.

Ma et al. (2021) proposed an automatic engagement recognition method that was based on the Neural Turing Machine. The model configuration is in two parts, namely, features extraction and features fusion. Features included the student's eye gaze, facial AUs, head pose, and body pose and were combined into a single final feature. Tests on the DAiSEE data set achieved an accuracy of 61.3 per cent.

Zhou et al. (2020) introduced a framework for the detection of emotional states using facial emotions in a learning set-up. The framework was based on a convolutional deep

neural network (CDNN). The CDNN classified individual emotions captured through webcam. Russel's model of core affects was employed in which emotions are categorised as four quadrants, namely, pleasant-active, pleasant-inactive, unpleasant-active and unpleasant-inactive. An accuracy of 66 per cent was reported using the VGG-S pretrained network which was tested on data collected from various sources.

Most of the previous works have not adequately incorporated a learning affects analysis into their models. Some works have only focused on the detection and classification of emotions. This research, therefore, proposes to contribute to the existing works. This is hoped to be achieved by including the analysis of learning affects from classified facial emotions and their application to online learning for estimating students' learning outcomes. This approach is expected to provide feedback to teachers and can lead to an improvement on the platform.

Methodology

The proposed research comprises the following stages:

- Stage 1: feature extraction using the pretrained ResNet-50, BiLSTM classifier for recognising emotions and mapping of emotions into positive, negative, and neutral learning affects.
- Stage 2: the model is tested with test data sets that have not been part of the training and validation, and further testing with a new data set is also completed.
- Stage 3: the model is tested with a real-life sample from the researcher's facial emotions and analysis of the learning affect.

Figure 1 shows the framework of the proposed model.

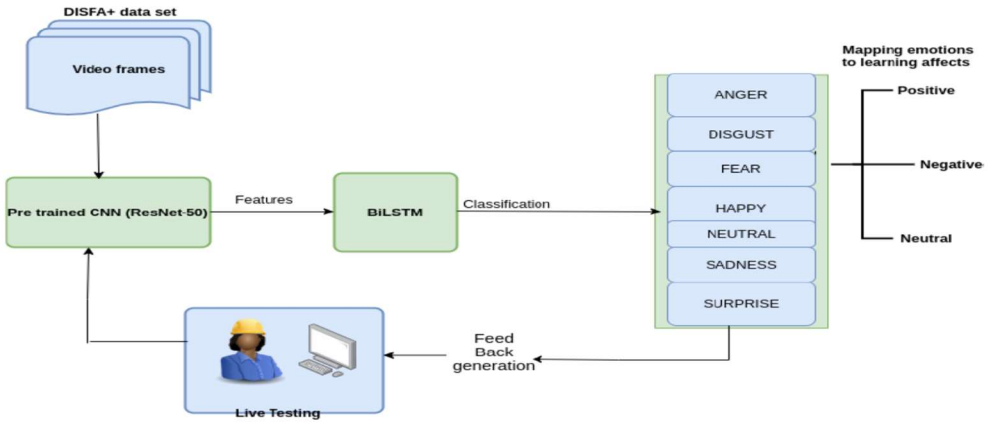


Figure 1: The proposed CNN-BiLSTM model

Data for Training and Testing

The DISFA+ data set is used for the experiment in this work. The DISFA+ is an extension of the DISFA data set (Mavadati et al. 2013).

The DISFA+ data set (Mavadati, Sanger, and Mahoor 2016) consists of videos and AU annotations of posed and spontaneous facial expressions from nine participants. AUs are the smallest visibly significant switch in facial action, which can be presented individually or in combinations to depict facial expressions (Li, Abtahi, and Zhu 2017). A high-definition (HD) camera was used to record the facial responses of the participants with 1 280 × 720 pixel resolution in 20 frames per seconds frame rate. The annotations include the intensity scores of 12 AUs for all of the frames. The DISFA+ data set is ground-truth data that comprises landmark points and subject-based self-reports. The DISFA+ data set is publicly available to researchers on request (Mavadati, Sanger, and Mahoor 2016). The experiment was conducted using a total of 26 603 frames from the video sequence. The work used 24 329 frames for the training and validation and 2 274 frames to test the model.

Feature Extraction

The features for the facial images of 24 329 frames for training and validation were extracted using the ResNet-50 pretrained network. The ResNet-50 network is a CNN with a depth of 50 layers, where the main idea is to skip blocks of convolution layers by using shortcut connections (Peng et al. 2020). Motivated by studies that proved that a deeper network is more powerful than a shallow network, the ResNet-50 was derived from 50-layer residual network architecture. The ResNet-50 was trained on the ImageNet database and can classify 100 categories of object (Peng et al. 2020). The pretrained ResNet-50 accepts the input video sequences which have been split into different frames of images. The frames were reshaped to 224 × 224 × 3 each, which is the input size required by the ResNet-50 network, added to a batch and passed through the ResNet-50 network for the extraction of discriminate features of the input images. The network produced a feature size of 24 329 × 2 048. The ResNet-50 has a 2 048 size of output dimensions.

Bidirectional LSTM

The bidirectional LSTM is an expansion of the LSTM model which is made up of two LSTMs. The first LSTM is applied to the input sequence data in the forward direction and the second LSTM takes in the reverse structure of the input sequence. This form of LSTM helps to improve learning long-term dependencies and can improve the accuracy of a model (Rahman et al. 2021). The forward hidden layer, h_t^f, processes the input in ascending order, that is t = 1, 2, 3 . . . T, whereas the backward hidden layer, h_t^b, processes the input sequence in descending order, that is t = T, . . . 3, 2, 1. Lastly, both layers are combined to generate output Y_t (Siami-Namini, Tavakoli, and Namin 2019).

The application of LSTM twice will lead to an improvement in learning long-term dependencies and, invariably, will improve the accuracy of the model (Baldi et al 1999). A BiLSTM is implemented using the following equations:

$$h_t^f = tanh\left(W_{xh}^f\, xt + W_{hh}^f\, h_{t-1}^f + b_h^f\right) \quad (1)$$

$$h_t^b = tanh\left(W_{xh}^b\, xt + W_{hh}^b\, hh_{t+1}^b + b_h^b\right) \quad (2)$$

$$Y_t = W_{hy}^f\, h_t^f + W_{hy}^b\, h_t^b + by \quad (3)$$

where
$tanh$ is an activation function of the hidden layer,
W is a weight matrix,
W_{xh} is a weight connecting input (x) to the hidden layer (h),
b is a bias vector for both the forward and backward hidden layers h_t^f and h_t^b in equations (1) and (2).

The extracted features were passed into the bidirectional LSTM for classification.

Evaluation Metrics

Evaluation metrics such as accuracy, precision, recall, and F-1 score were used to evaluate the proposed model. The computations for precision (P), recall (R) and F-1 score are given as follows:

$$Precision\ (P) = \frac{TP}{TP+F} \quad (4)$$

$$Recall\ (R) = \frac{TP}{TP+FN} \quad (5)$$

$$F-1Score = \frac{2*P*R}{P+R} \quad (6)$$

where
TP denotes True Positive,
TN denotes True Negative,
FP denotes False Positive,
FN denotes False Negative.

Experiment Details

The implementation was carried out using a CPU Intel Core i7, 8th generation, with 16GB RAM and Linux operating system. The codes were run in a tensor-flow environment, with a tensor-board to visualise the training epoch accuracy and loss. The

data set was split into training, validation and testing data sets. The work stages included training and validation and mapping of recognised emotions.

Training and Validation

The extracted features of size 24 329 × 2 048 were split into a training data set at the ratio of 75:25 of size 18 246 × 2 048 samples and validation data sets of size 6 083 × 2 048. They were then passed to the classifier, using a batch size of 128 and the maximum epochs set to 100. Binary cross entropy was used to identify the loss function and the sigmoid function was used to compute the probability of the output layer. Early stopping was used to avoid overfitting and the training stopped at the 61st epoch. The performance metric that was used for the learning accuracy is the F-1 score and confusion matrix. The BiLSTM network classifies the input features of size 18 246 × 2 048; both the temporal and spatial features of the samples are considered by the LSTM and validate the classification using the validation data sets of size 6 083 × 2 048 and gives an accuracy over the validation data set. The network was also tested using the remaining part of the DISFA+ data set that has not been exposed to the network before and the accuracy was also recorded. The classified emotions include anger, disgust, fear, happiness, sadness, surprise and neutral.

Mapping of Recognised Emotions to Learning Affect

In this study, we explored various literature to map the classified emotions of anger, disgust, fear, happiness, sadness, surprise and neutral into learning affects. The work adopts the mapping of Sathik and Jonathan (2013), Kapoor et al. (2001), Pan, Wang and Luo (2018), and Zakka and Vadapalli (2020). These works were used to decide the direct mapping between facial emotions and learning affect in this work. Table 1 provides the summary of the results of the mapping of the various emotions based on literature for the study.

Table 1: Mapping of emotions to learning affects

Emotion	Sathik and Jonathan	Kapoor et al.	Pan, Wang and Luo	Zakka and Vadapalli	Current Study
Anger	Negative	Negative	Positive	Negative	Negative
Disgust	–	Negative	Negative	Negative	Negative
Fear	Positive	Positive	–	Positive	Undefined
Happiness	–	Positive	Positive	Positive	Positive
Sadness	Negative	Negative	–	Negative	Negative
Surprise	Positive	Positive	Positive	Positive	Positive
Neutral	–	–	Positive	Positive	Neutral

Results

Testing

The model was tested with 2 274 samples of the DISFA+ data set which had not been exposed to the model before. Figure 2 illustrates the confusion matrix of the test result of the CNN-BiLSTM classification model.

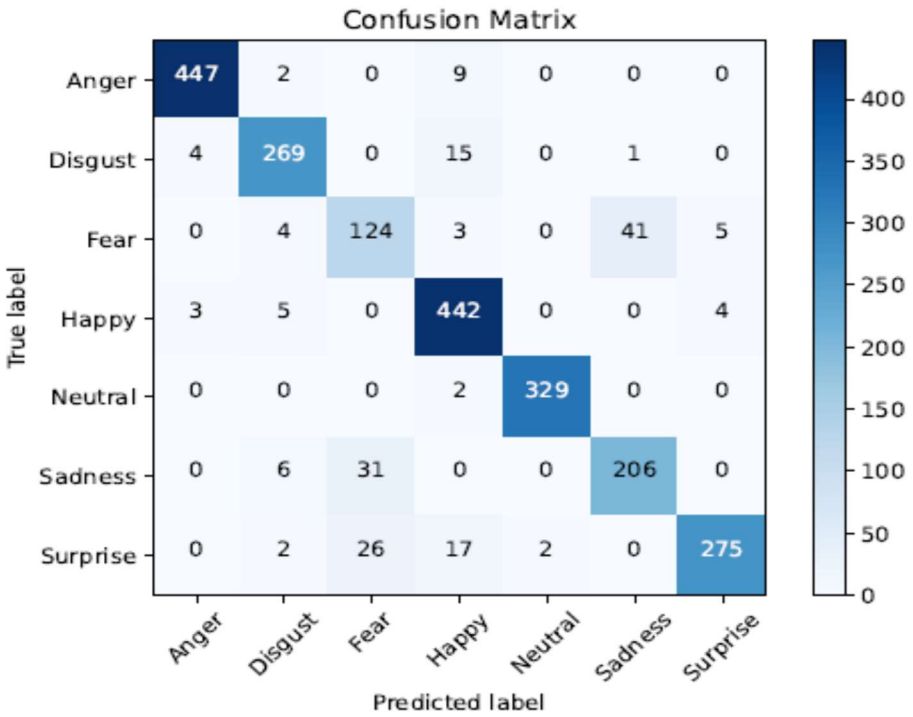


Figure 2: The confusion matrix of the test result

The CNN-BiLSTM model achieved a 92 per cent test accuracy. Obvious instances of emotions that were incorrectly classified include 2.5 per cent of anger, 7.4 per cent of disgust, 42 per cent of fear, 2.7 per cent of happiness, 0.6 per cent of neutral, 18 per cent of sadness, and 17 per cent of surprise emotions. Table 2 presents the summary of the test results. The detected emotions were then mapped into positive, negative, and neutral learning affects based on literature as presented in Table 1.

Table 2: Summary of the test results

Emotion	Precision	Recall	F-1 Score	Number of sample
Anger	0.98	0.98	0.98	458
Disgust	0.93	0.93	0.93	289
Fear	0.69	0.70	0.69	177
Happiness	0.91	0.97	0.94	454
Neutral	0.99	0.99	0.99	331
Sadness	0.83	0.85	0.84	243
Surprise	0.97	0.85	0.91	322
Accuracy			0.92	2 274

The work explored the mapping further through another test. Gupta et al. (2016) opined that the seven basic expressions might not be enough to estimate the learning affects for a prolonged learning condition, because they are liable to rapid changes. They therefore compiled a data set called the Dataset for Affective States in E-learning Environments (DAiSEE). The DAiSEE was labelled with the affective states of boredom, frustration, engagement and confusion. The study therefore tested the model with the DAiSEE data set. This research performed an experiment by testing the proposed model with 2 727 unlabelled samples from the DAiSEE data set. From the results, the study was able to compare the true labels of the samples with the predicted labels by the proposed model. The results are presented in Table 3.

Table 3: Experiments with DAiSEE samples and the predicted output

DAiSEE Labels	Total Samples Tested	Classification by CNN-BiLSTM					
		Anger	Disgust	Fear	Happiness	Sadness	Surprise
Engagement	691	0	0	0	354	0	337
Boredom	683	0	683	0	0	0	0
Confusion	923	0	0	390	0	533	0
Frustration	430	430	0	0	0	0	0
Total samples tested = 2 727							

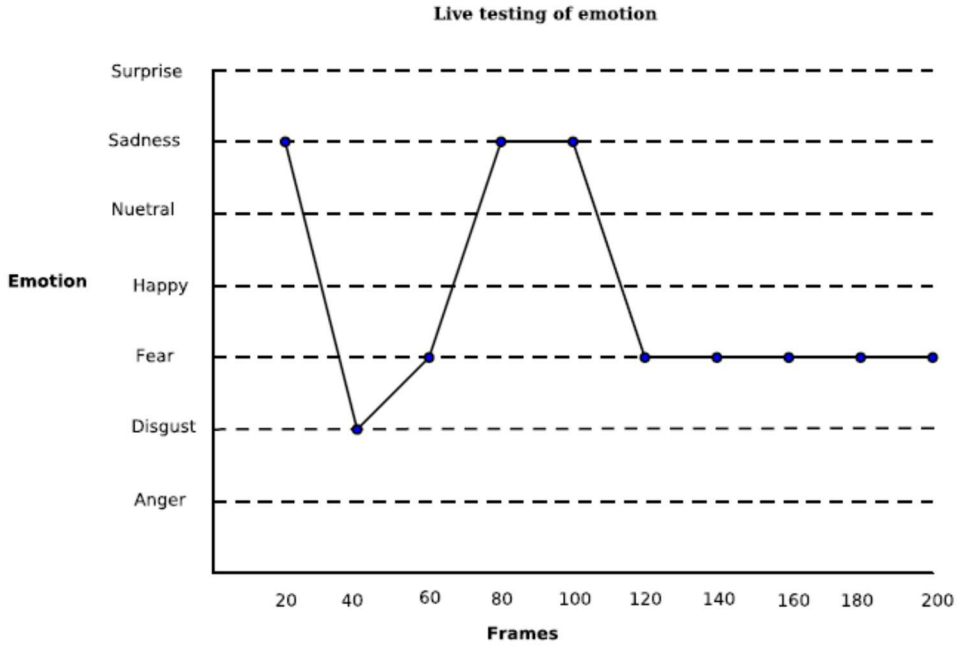
From the results of the experiment, it is clear that the happiness and surprise emotions are mapped to the engaged affective states, which indicates positive learning affects. The fear and sadness emotions were mapped to the confused affective states, which indicates negative learning affects. The disgust emotion is mapped to the boredom affective states, which indicates a negative learning affect. Anger is mapped to the frustrated affective states, which also indicates a negative learning affect. The work therefore verifies the previous mapping in Table 1. The research therefore infer that the seven basic emotions that was trained in the work can also be mapped to these affective

states to estimate the learning affects. Furthermore, based on the mapping of fear in the experiment, the work suggests that fear be a negative learning affect against the results by Sathik and Jonathan (2013), Kapoor et al. (2001), and Zakka and Vadapalli (2020).

Live Testing

The work further tested the model using a single data sample from the researcher to determine how well the model could classify facial emotion expression in real life and to analyse the learning affect estimation through the mapping of emotions by watching a video for 10 minutes. After the video processing by the proposed model, the model came up with a graph as shown in Figure 3. The graph shows the changes of emotions which are plotted at different frames within 10 minutes. The live testing was necessary for verifying the accuracy of the model and the validity of the various mappings.

Figure 3: The graph of emotion changes with frames during live testing



Discussion

The work deduced from the experiments that it is possible to estimate the students' learning affects using the facial emotion expression recognition, mapping the emotions classified to positive, negative and neutral learning affects, and analysing the learning affects from the mapping. The emotions can also be mapped to the affective states of boredom, engagement, confusion and frustration. From the model testing, the system has a probability of an eight per cent tendency of misclassifying emotions, with the fear emotion having the highest misclassification. Generally, the probability of the affect analysis being wrong is very low. From the live testing, the researcher affirms that facial

emotion expression is an effective way of estimating the students' learning affects in an e-learning platform.

Conclusion and Future Works

This work has presented an area of research in machine learning which seeks to deal with human and computer interaction in the education system especially during the pandemic. It is aimed at supporting teachers during online classes, which is the most common and safest teaching and learning method adopted during the prevailing present condition. The study combined the use of the machine learning models by using the DISFA+ data set and analysis of learning affects and applied them to the online learning platform. The study achieved a classification accuracy of 92 per cent on the test data set. Various deep-learning approaches have been adopted using the DISFA+ data set. The approach used in this work has achieved a comparable performance to other state-of-the-art methods. The work tested the model using another data set which comprised the affects emotions, called the DAiSEE data set. The work was able to infer that the seven basic emotions can also be mapped to the affective emotions. The study concludes that analysing the changes of the facial emotions exhibited by students in online learning and mapping their emotions to various learning affects can help to determine the learning outcomes of students.

It is desired that this work improve the online platform and, hopefully, be able to deal with the current realities and to tackle envisaged future challenges. It is desired that this model be improved in the near future, especially since more features such as reasoning about the learning affects can be included in the model. Future researchers can also use other approaches to affects analysis.

References

Ayvaz, Uğur, Hüseyin Gürüler, and Mehmet Osman Devrim. 2017. "Use of Facial Emotion Recognition in E-Learning Systems." *Information Technologies and Learning Tools* 60 (4): 95–104. https://doi.org/10.33407/itlt.v60i4.1743.

Baldi, Pierre, Søren Brunak, Paolo Frasconi, Giovanni Soda, and Gianluca Pollastri. 1999. "Exploiting the Past and the Future in Protein Secondary Structure Prediction." *Bioinformatics* 15 (11): 937–46. https://doi.org/10.1093/bioinformatics/15.11.937.

De Carolis, Berardina, Francesca D'Errico, Nicola Macchiarulo, Marinella Paciello, and Giuseppe Palestra. 2020. "Recognizing Cognitive Emotions in E-Learning Environment." In *International Workshop on Higher Education Learning Methodologies and Technologies Online*, edited by L. S. Agrati, 17–27. Cham: Springer. https://doi.org/10.1007/978-3-030-67435-9_2.

Dewan, M. Ali Akber, Mahbub Murshed, and Fuhua Lin. 2019. "Engagement Detection in Online Learning: A Review." *Smart Learning Environments* 6 (1): 1–20. https://doi.org/10.1186/s40561-018-0080-z.

Dhankhar, Poonam. 2019. "ResNet-50 and VGG-16 for Recognizing Facial Emotions." *International Journal of Innovations in Engineering and Technology* 13 (4): 126–30.

Ferdig, Richard E., Emily Baumgartner, Richard Hartshorne, Regina Kaplan-Rakowski, and Chrystalla Mouza. 2020. *Teaching, Technology, and Teacher Education during the COVID-19 Pandemic: Stories from the Field*. Association for the Advancement of Computing in Education Waynesville.

Gupta, Abhay, Arjun D'Cunha, Kamal Awasthi, and Vineeth Balasubramanian. 2016. "DAiSEE: Towards User Engagement recognition in the wild. arXiv preprint arXiv:1609.01885.

Kapoor, Ashish, Selene Mota, Rosalind W. Picard, et al. 2001. "Towards a Learning Companion that Recognizes Affect." In *AAAI Fall symposium, volume 543*, 2–4.

Khanzada, Amil, Charles Bai, and Ferhat Turker Celepcikay. 2020. "Facial Expression Recognition with Deep Learning." Cornell University. https://arxiv.org/abs/2004.11823.

Klein, Richard, and Turgay Celik. 2017. "The Wits Intelligent Teaching System: Detecting Student Engagement during Lectures Using Convolutional Neural Networks." In *2017 IEEE international conference on image processing (ICIP)*, 2856–60. IEEE. https://doi.org/10.1109/ICIP.2017.8296804.

Li, Wei, Farnaz Abtahi, and Zhigang Zhu. 2017. "Action Unit Detection with Region Adaptation, Multi-Labeling Learning and Optimal Temporal Fusing." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1841–50. https://doi.org/10.1109/CVPR.2017.716.

Ma, Xiaoyang, Min Xu, Yao Dong, and Zhong Sun. 2021. "Automatic Student Engagement in Online Learning Environment Based on Neural Turing Machine." *International Journal of Information and Education Technology* 11 (3): 107–11. https://doi.org/10.18178/ijiet.2021.11.3.1497.

Marcelino, P. 2018. "Transfer Learning from Pre-Trained Models." *Towards Data Science*. Accessed 8 October 2021. https://towardsdatascience.com/transfer-learning-from-pre-trained-models-f2393f124751.

Mavadati, S. Mohammad, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. 2013. "DISFA: A Spontaneous Facial Action Intensity Database." *IEEE Transactions on Affective Computing* 4 (2): 151–60. https://doi.org/10.1109/T-AFFC.2013.4.

Mavadati, Mohammad, Peyten Sanger, and Mohammad H. Mahoor. 2016. "Extended DISFA Dataset: Investigating Posed and Spontaneous Facial Expressions." In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 1–8. IEEE. https://doi.org/10.1109/CVPRW.2016.182.

Megahed, Mohammed, and Ammar Mohammed. 2020. "Modeling Adaptive E-Learning Environment Using Facial Expressions and Fuzzy Logic." *Expert Systems with Applications* 157: 113460. https://doi.org/10.1016/j.eswa.2020.113460.

Mukhopadhyay, Moutan, Saurabh Pal, Anand Nayyar, Pijush Kanti Dutta Pramanik, Niloy Dasgupta, and Prasenjit Choudhury. 2020. "Facial Emotion Detection to Assess Learner's State of Mind in an Online Learning System." In *Proceedings of the 2020 5th International Conference on Intelligent Information Technology*, 107–115. https://doi.org/10.1145/3385209.3385231.

Pan, Minyu, Jing Wang, and Zuying Luo. 2018. "Modelling Study on Learning Affects for Classroom Teaching/Learning Auto-Evaluation." *Science* 6 (3): 81–86. https://doi.org/10.11648/j.sjedu.20180603.12.

Pekrun, Reinhard. 1992. “The Impact of Emotions on Learning and Achievement: Towards a Theory of Cognitive/Motivational Mediators.” *Applied Psychology* 41 (4): 359–76. https://doi.org/10.1111/j.1464-0597.1992.tb00712.x.

Peng, Jie, Shuai Kang, Zhengyuan Ning, Hangxia Deng, Jingxian Shen, Yikai Xu, Jing Zhang, Wei Zhao, Xinling Li, Wuxing Gong, et al. 2020. “Residual Convolutional Neural Network for Predicting Response of Transarterial Chemoembolization in Hepatocellular Carcinoma from CT Imaging.” *European Radiology* 30 (1): 413–24. https://doi.org/10.1007/s00330-019-06318-1.

Phillips, Jennifer. 1993. “Nonverbal Communication: An Essential Skill in the Workplace.” *Australian Medical Record Journal* 23 (4): 132–34. https://doi.org/10.1177/183335839302300406.

Pise, Anil, Hima Vadapalli, and Ian Sanders. 2020. “Facial Emotion Recognition Using Temporal Relational Network: An Application to E-Learning.” *Multimedia Tools and Applications* 1–21. https://doi.org/10.1007/s11042-020-10133-y.

Power, Michael J., and Tim Dalgleish. 1999. *Handbook of Cognition and Emotion*. Chicester: Wiley.

Rahman, M. D., Yutaka Watanobe, Keita Nakamura, et al. 2021. “A Bidirectional LSTM Language Model for Code Evaluation and Repair.” *Symmetry* 13 (2): 247. https://doi.org/10.3390/sym13020247.

Ray, Arindam, and Amlan Chakrabarti. 2012. “Design and Implementation of Affective E-Learning Strategy Based on Facial Emotion Recognition.” In *Proceedings of the International Conference on Information Systems Design and Intelligent Applications*, Visakhapatnam, India, January 2012, 613–622. Springer. https://doi.org/10.1007/978-3-642-27443-5_71.

Sathik, Mohamed, and Sofia G. Jonathan. 2013. “Effect of Facial Expressions on Student’s Comprehension Recognition in Virtual Educational Environments.” *SpringerPlus* 2 (1): 1–9. https://doi.org/10.1186/2193-1801-2-455.

Siami-Namini, Sima, Neda Tavakoli, and Akbar Siami Namin. 2019. “The Performance of LSTM and BiLSTM in Forecasting Time Series.” In *2019 IEEE International Conference on Big Data*, 3285–3292. IEEE. https://doi.org/10.1109/BigData47090.2019.9005997.

Spector, J. Michael, M. David Merrill, Jan Elen, and Martin J. Bishop. 2014. *Handbook of Research on Educational Communications and Technology*. London: Springer. https://doi.org/10.1007/978-1-4614-3185-5.

Sun, Ai, Ying-Jian Li, Yueh-Min Huang, and Qiong Li. 2017. “Using Facial Expression to Detect Emotion in E-Learning System: A Deep Learning Method.” In *International Symposium on Emerging Technologies for Education*, edited by T. C. Hunang, R. Lau, Y. M. Huang, M. Spaniol and C. H. Yen, 446–455. Cham: Springer. https://doi.org/10.1007/978-3-319-71084-6_52.

Tyng, Chai M., Hafeez U. Amin, Mohamad N. M. Saad, and Aamir S. Malik. 2017. “The Influences of Emotion on Learning and Memory.” *Frontiers in Psychology* 1454–60. https://doi.org/10.3389/fpsyg.2017.01454.

Wahab, Noorul, Asifullah Khan, and Yeon Soo Lee. 2019. “Transfer Learning Based Deep CNN for Segmentation and Detection of Mitoses in Breast Cancer Histopathological Images.” *Microscopy* 68 (3): 216–33. https://doi.org/10.1093/jmicro/dfz002.

Zakka, Benisemeni Esther, and Hima Vadapalli. 2020. “Estimating Student Learning Affect Using Facial Emotions.” In *2020 2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, 1–6. IEEE. https://doi.org/10.1109/IMITEC50163.2020.9334075.

Zhou, Wenbin, Justin Cheng, Xingyu Lei, Bedrich Benes, and Nicoletta Adamo. 2020. “Deep Learning-Based Emotion Recognition from Real-Time Videos.” In *International Conference on Human–Computer Interaction*, edited by M. Kurosu, 321–332. Cham: Springer. https://doi.org/10.1007/978-3-030-49062-1_22.

Quality Impact of Accommodating Customer Requirements through Plug-Ins and Configuration Files

Geoffrey Lydall
https://orcid.org/0000-0001-9461-9290
School of Electrical and Information Engineering, University of the Witwatersrand, South Africa
geoffrey.lydall@gmail.com

Stephen Phillip Levitt
https://orcid.org/0000-0001-6054-6134
School of Electrical and Information Engineering, University of the Witwatersrand, South Africa
stephen.levitt@wits.ac.za

Abstract

A case study is conducted on a weighbridge application which has been modified to accommodate the requirements of three different customers. The application architecture allows for different configurations and customer-specific modifications. The impact of customisations and modifications on the structural and functional quality of the system is evaluated. The structural quality is measured using the Maintainability Index and other metrics. The functional quality is assessed using defect data recorded in the task-tracking software Jira. The amount of modification is measured using the number of rules defined per customer. The results indicate that structural quality is unaffected by the modifications and that the functional quality is reduced as more customisation rules are defined, indicating a partial success of the architecture.

CCS concepts: software and its engineering, designing software, software architectures
Keywords: software quality metrics, plug-in architecture, customisation

Introduction

Software, from individual applications to systems for enterprise resource planning (ERP), can often be customised by allowing the base functionality to be extended to support differing business requirements for a variety of customers.

Software vendors achieve this by making use of configurable architectures which are designed to accommodate certain customisations or modifications that are requested by customers. One approach to creating a configurable architecture is to use a dependency injection framework (Razina and Janzen 2007) that can be configured to override the original class types used at runtime with types defined in custom assemblies. In this way, all customers share common basic functionality and the software is extended polymorphically with new behaviours that fulfil the new business requirements.

The aim of using a configurable architecture is to reduce the effort in accommodating varying customer requirements. However, customisations could lead to problems of reduced software quality and increase the cost of maintenance. Software quality can be considered along three principal dimensions: process, structural, and functional quality (Chappell 2011). Customising software should not affect the development process and methodology but it might have an impact on both the structural and functional quality of the system.

A configurable architecture can therefore be regarded as successful from a quality perspective if it exhibits the following qualities:

- The architecture is able to accommodate changes without compromising the maintainability (internal or structural quality) of the software.
- The architecture provides developers with enough freedom to effect functional changes in the software to meet differing business requirements. This means that the architecture does not impede the development team's ability to correctly express custom business requirements or business rules. The team is able to do so without introducing unintended side effects and affecting the external or functional quality.

Given the above two premises, a case study is presented here which attempts to answer the following research questions:

- In the case of a system that is designed for configurability, how does the structural quality of customisations compare with the rest of the system?
- In the case of a system that is designed for configurability, what effect does implementing customisations have on the functional quality as perceived by the user?
- In the context of the previous two questions, how successful is the architecture in supporting the changes required by different customers?

The case study presented here centres on a weighbridge application which supports customisation through dependency injection. Software metrics are used for both quantifying the customisations made to the weighbridge application in order to meet the demands of different customers and measuring the effect that these changes have on the quality of this application. This study does not investigate concerns related to the customisations such as why the customer's business processes have the requirements that they do, the quality of the specifications regarding the custom requirements, and the skill level of the teams performing the customisations.

Such concerns are important in understanding quality; however, the focus here is to determine whether customising software reduces quality or not, as opposed to determining the root causes of any quality issues that arise.

Answers to the above research questions can provide insight into how successful a dependency injection architecture is at accommodating the specific business needs of a customer from a quality perspective. The findings of this study serve to increase the body of knowledge that can inform business regarding build (develop a bespoke software solution) versus "buy-and-customise" decisions for new software systems.

Related Work

ERP Customisation

ERP systems have been adopted in business and are at times customised to meet the specific needs of a customer in the workplace (Light 2001). Light's work discusses the maintenance implications of various changes to an ERP system for both the customer and the vendor. Some of the challenges include a customisation for one customer which actually competes with the customisations for another customer. Light also points out the fragility in the upgrade path of customisations.

Light's work offers insight into the types of change that have been applied to ERP systems for customers. The following types of change in customising the software (in order of potential for required maintenance, high to low) are noted:

- change functionality;
- add functionality;
- process automation;
- amend reports or displays; and
- new reports.

Guido and Pierluigi (2008) assess the feasibility of ERP implementation extensions. This study provides a formal understanding of the business context in which a system is going to operate. They specifically mention the alignment of the ERP system to the business. Guido and Pierluigi's work has been applied in numerous studies, such as in the study by Kumar, Suresh and Prashanth (2009). They conclude that customisation beyond 30 per cent adds considerable risk to the project. The work outlines customisations as "any modifications or extensions that change how the out-of-box ERP system works" (Kumar, Suresh, and Prashanth 2009). Their conclusion stems from an analysis of error counts in modules of the system and recommendations from the system vendor.

Functional Quality

Potential metrics for measuring functional quality include the goal-question-metric (GQM) approach (Basili and Rombach 1988; Fenton and Pfleeger 1998), which provides a measure of functional usefulness or fit-for-purpose metrics. Hall and Fenton's (1997) work furthers this approach. Recent refinements to this approach

include the work by Kelemen, Bényász and Badinka (2014), which proposes the term "measurement based software quality assurance framework". At the core of these approaches is quantifying how fit for purpose a system is. These approaches are useful when assessing the value of software customisations and require a survey to be conducted involving the users of the systems being measured.

Alternatively, functional quality can be inferred from the absence of defects and the number of reported defects can be measured. Freimut, Denger and Ketterer's work (2005) considers approaches for measuring defects for the purposes of quality improvement extensions. This work provides a process to track defect introduction and detection for establishing a quality assurance baseline.

Case Description and Approach

A case study is conducted into a weighbridge application that has been deployed to numerous customers in varying industries. From Fenton and Pfleeger (1998), it is appreciated that case studies are difficult to control and reproduce; however, the strength of a case study is that the software has been produced entirely outside the control of the researchers. It therefore presents a completely realistic scenario (as opposed to a laboratory or field experiment).

Given the work from Easterbrook et al. (2008), the research question can be said to be of a causal nature. The case study is to be used in an exploratory manner to investigate the phenomena of what effect the modification of software has on its internal and external quality attributes.

Three customers of the weighbridge application have been selected for the nature of their modifications and the researchers' perception of the customers. The data gathered in the case study is used to answer the following questions:

- How many bugs have been logged per customer?
- What is different about the software deployed for each customer?
- By how much does the software configuration and customisation differ between each customer?
- How does the amount to which the requirements for these customers differ relate to the number of bugs introduced for these customers?
- How does the number of issues raised against a particular customer relate to other code metrics for that customer?

The number of bugs logged per customer will be used as the measure of quality for the software as experienced by the customer. One potential source of error in this methodology is that it assumes that the same process is applied by all customers when

a defect is encountered. For example, the willingness to log a defect or the ability to identify or describe a defect for each customer may not be the same.

In order to determine the way in which the amount of customisation relates to the number of bugs introduced, a means for measuring the amount of change is needed. This requires an understanding of the selected customers and the way in which the software is changed to meet their requirements.

The Weighbridge Application

The weighbridge application is designed to manage the weigh-in and weigh-out of freight in order to verify that the load of a shipment that was dispatched matches the load received at the destination. A major component of the product offering is integration into existing systems for ERP. These integrations are not considered in the scope of this study owing to limited access to data pertaining to the integrations.

A shipment is the central concern of the weighbridge application domain. The shipment is contained within a dispatch transaction. The dispatch transaction contains information about periphery concerns such as the vehicles participating in the shipment, vehicle drivers, and the description and mass of the payload. Other examples of information that can be tracked include legislative compliance information, unique identifiers, and custom fields which can be configured to store arbitrary data.

The application has been in operation for several years in the mining and agricultural industries and, more recently, it has been used for road ordinance (checking for overweight vehicles). The application is in its third major revision since its initial release. The application is written in C# and uses Microsoft SQL Server for data persistence. Revision control was previously managed by SVN (Apache Software Foundation 2018); however, this has been migrated to Git. A full revision history is available.

As of its third revision (the revision under study), the application is designed to accommodate a series of customisations by way of features that can be configured. The supported customisations are diverse, ranging from branding and user interface (UI) label changes to custom process definitions, data fields, validations, and behaviours. The configurable features include enabling and disabling visual controls; specifying validation rules and custom commands on actions (for example, do “x” on save).

Customising the Application

In order to allow customisation, such as custom commands on save, the application features an architecture that can dynamically load plug-in libraries, as specified by configuration files. These files are defined per customer. Certain kinds of configuration simply specify options for built-in components, whereas other configurations use code injection through a compiled plug-in assembly. Configuration options that are “built-

in" are referred to as core configurations, whereas the configurations that are sourced from plug-ins are referred to as customised configurations.

Configurable behaviours are facilitated by dependency injection (Fowler 2004) and the model-view-viewmodel (MVVM) (Smith 2009) pattern.

The plug-in architecture provides direct persistence for custom data where required. This is achieved through the use of data bags and generic data fields that can be used by a custom plug-in if required. Should generic data fields not be sufficient to accommodate a customer requirement, then the core code of the application is extended to accommodate the new requirement.

The application is customised for each customer by specifying configurations for the application. Configurations are specified in XML files (see Figure 1). The configuration files contain a collection of rule contexts that are identified by a key in order that the application modules can identify the appropriate rule contexts to configure the module. The rule contexts can contain a view-model, a view, a model, and other rule contexts. The rule context itself is a container for a module's configuration and a single rule context counts as a rule.

A behaviour change is achieved by specifying extensions in the view-model portion of the configuration. Extensions will be executed from particular UI events (for example, CanSave, Save). Extensions are injected from either core assemblies (built-ins) or a customer-specific assembly, which is referred to as a plug-in, as defined in the view-model portion of the configuration rule context.

Visual changes can be driven either by configuring either the view-model or the view itself. Behavioural (for example, validation on or off) UI changes are typically achieved by specifying a view-model setting for the view-model in the rule context. Text labels and control visibility are configured by specifying a view-control setting in the view portion of the configuration rule context.

For example, a form such as indicated in Figure 1 can be customised in a number of ways. View-control settings could be used to hide the "Horse" field (a horse is the part of a truck with the engine and cabin that pulls the trailer) or to change the text label for "CustomField1" to "Permit". View-model settings can be used to set the validation rules for the country weight limits or to inject a custom behaviour for disabling the submit button if, for example, the truck exceeds the legal weight limit. The ACME Corporation, for example, can inject a specific ERP integration into its organisation upon submitting using a pre-save action. A custom task can also be performed as a post-save action when submitting (for example, printing a weighbridge ticket).

In summary, the application is customised for a customer by defining rules which can

- set configurable options within the application, and
- inject custom code that will be executed on specific actions.

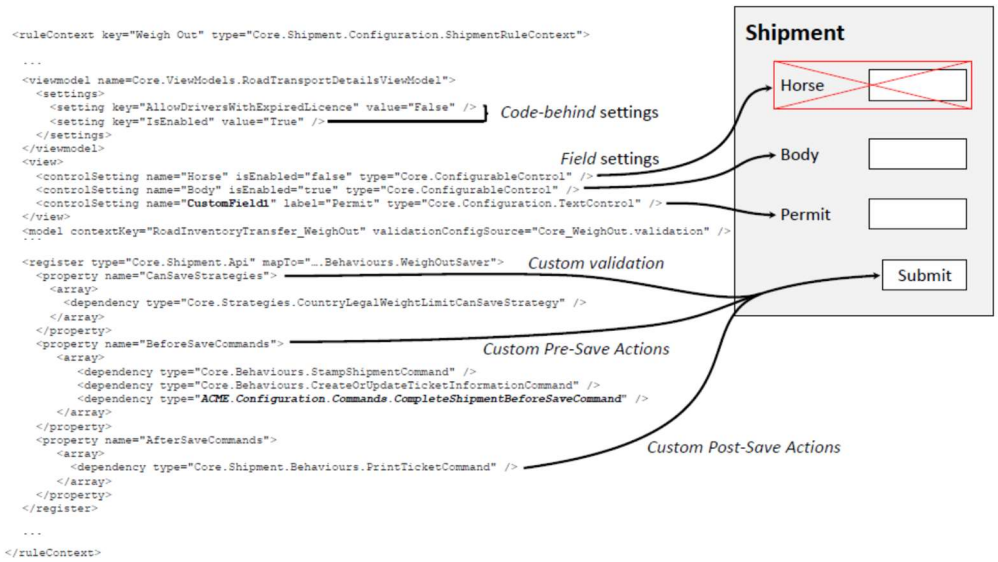


Figure 1: Example form for customisation

Customer Selection and Degrees of Customisation

Only customers using a recent version of the software are considered to ensure that an appropriate comparison is being made.

The degree of customisation was considered when selecting customers for the study to ensure that the selected customers exhibit different levels of customisation. These levels are:

- out-of-the-box (little to no customisation);
- heavy customisations; and
- domain modified.

C1 – Out-of-the-Box Customisations

The out-of-the-box customer is one requiring little to no customisation of the product, with at most a configuration of the built-in feature controls. This implies that any "custom" code for this customer should then lie mostly within the realm of configuration.

Customer C1, which operates in the agricultural sector and deals with shipments of agricultural produce, was selected to represent this category.

C2 – Heavy Customisations

A customer with heavy customisations or integrations is one where the changes include a specific addition or change of features. These changes are presented with a higher number of rules defined for the customer than for C1, and they include plug-in code specified in the configuration.

Customer C2, which operates in the mining sector and deals with shipments of ores, was chosen to represent this category.

C3 – Domain Modified

A customer which constitutes having a domain-modified classification means that the customer's requirements introduce a change to the software which the original domain or design could not accommodate. These changes are therefore accommodated by introducing new domain concepts into the software – in other words, this falls outside the realm of the configuration architecture.

Customer C3 is a roads development agency which operates a national road network and uses the software to record and assist with enforcing compliance with the vehicle weight limit.

This customer required a different process compared to the process originally supported by the software. The original process was for a truck, loaded with goods to "weigh-out" from a location with a specific mass, and for the same truck to "weigh-in" at its destination and for the masses to be compared and checked that they match. Customer C3 required changes to this model because it requires that all trucks only ever weigh-in. Furthermore, in the original model, a single vehicle's data life cycle is limited to a single shipment, whereas customer C3 requires that the truck be added to a running account to which charges can be levied and tracked.

Threats to Validity

A number of threats to validity (Easterbrook et al. 2008) are identified. The largest threat to validity is an external threat, which is presented by a case study consisting of a single product across only three customers. The data presented for these cases may not represent a general population of customers and software systems.

An internal threat to validity occurs when the full nature of variability between customers cannot be completely known. For example, quality is measured by defect and therefore assumes that defects are logged to each customer using an equivalent process. This can only be known through greater engagement with the software life cycle, including the vendor and customer staff. This may be offset by including more cases in the study.

Another potential internal threat to validity occurs when the collected data treats all counts uniformly. For example, all rules and bugs are treated equally although some rules will have more impact than others and some bugs will be more severe than others.

Metrics

The amount of customisation can be quantified on a number of dimensions, including the number of defined configurations for the customer (both core and customised) and how much code has been written for any customisations.

Given that each customer's installation is customised by way of configuration files, the amount of customisation can be measured by the number of configured items for the customer. These items can be further classified as either built-in options or custom plug-ins.

Measurements pertaining to customisation can be sourced from a variety of artefacts. The data sources that have been used in this study are the following:

- source code;
- configuration files; and
- revision history and issues list.

It is worth noting that although the business requirement specifications reflect the customisations required and may seem useful, a focus on metric extraction from the source code and configuration files has been preferred. This is because the configuration files are XML formatted making them easier to parse and analyse than the free-text specification documents. The configuration files also use very specific and consistent terms (since they configure the application) which make natural grouping categories for analysis.

The following measurements are extracted from the above data sources:

4 Source code
   – Maintainability index (MI) – per module
      * Lines of code (LOC)
      * Cyclomatic complexity
      * Count of methods
      * Count of calls to other classes
      * Count of fanout
5 Configuration files
   – Count of total rules

- Count of rules configured to use custom code
- Count of core configurations unique to customer
- Count of customised configurations

6 Jira

- Count of issues per customer

The quality is determined in two dimensions: A measured set of code metrics and the measured number of defects. This allows the determination of structural quality (from the code metrics) and functional quality (from the defects). A success or failure is then determined by comparing the resulting qualities for varying degrees of customisation. For both structural and functional quality, the architecture can be considered a success if they are unaffected by the customisation.

Structural Quality Metrics

Some of the desired metrics can be acquired using Visual Studio tooling. In particular, the FxCop (Kresowaty 2008; Microsoft Contributors 2018) metrics calculator was used on the relevant assemblies to produce XML files which contain metrics for the assembly. FxCop will produce the following metrics at varying levels of detail (assembly; namespace; type or class; member or method):

- LOC;
- depth of inheritance;
- cyclomatic complexity;
- class coupling; and
- MI.

The metrics of LOC and depth of inheritance are self-describing. LOC is not a measure of the absolute LOC, but rather a measure of the size of code that lives inside methods, i.e. ignoring class definitions, declarations and other language features. Specifically, Visual Studio will perform an estimated count of the number of LOC based on the compiled common intermediate language (CIL). Although this will not include class definitions or declarations, it will include "invisible" code such as default constructors and initialisers.

The cyclomatic complexity is best described as the number of branches in a program, including method calls. Visual Studio will count an interface as a class when determining the number of classes. This is because the metrics tool uses the CIL to perform the measurements, and an interface is internally represented as a class. These interfaces have a cyclomatic complexity of one per method since each method represents a branch in code. Other language features such as getters, setters, and default constructors will also introduce a cyclomatic complexity of one per language feature.

The class coupling metric refers to the number of classes that the unit under measure collaborates with, i.e. the number of distinctly referenced classes in the code. Both class and interface types constitute a coupling. The class coupling metric, Fanout, counts all references, inheritance hierarchies, and type checks on fields. Type checks on properties, however, will not result in an increased coupling because the calling method is decoupled from the actual type by the getter method. Inheritance hierarchies are counted because the derived class's constructor must also make a call to the constructor of the base class. This does not happen for interfaces because interfaces do not have constructors. Each getter or setter will also indicate a coupling of one if it is of another type. Visual Studio will also count coupling for library types (for example, System.Console). However, the coupling is rolled up in a namespace or module according to the number of unique classes in that scope, which means that the tool uses the total sum of coupling for each class.

The MI metric (discussed in detail below) provides an indication of how maintainable an application is.

Maintainability Index

The extraction of established code metrics such as the MI provides an objective measure of the way in which the internal, structural quality of the application has changed. An MI measurement can be performed using Microsoft Visual Studio (Microsoft Corporation 2018), which is significant as the application is written using .NET technologies and therefore making MI the preferred metric. The MI measurement is captured for both modifications and additions allowing for a further comparison of the effect of each on quality. In other words, it can then be determined whether the nature of a change (addition or modification) has any impact on structural quality.

Of note is that the MI as produced by Visual Studio is not identical to the Software Engineering Institute promoted index introduced by Oman and Hagemeister in 1992 but has two minor differences (Van Deursen 2014). The first is that Oman and Hagemeisters' (1994) original index had a range of 0 to 171. The Visual Studio Team has normalised this metric to the range [0, 100]. Secondly, the original index also included a factor for the number of comments in the code (Coleman et al. 1994), but this is not included in the Visual Studio metric. One motivation for excluding comments from the metric could be the in-line XML Code Doc in C# (that lives in the code files) which would create an unusually large number of comments per LOC creating an apparent increase in code maintainability as created by documentation rather than comments.

The index itself is the result of a study by Oman and Hagemeister (1994) which involved a regression analysis on several software systems written in C and Pascal. A range of metrics were gathered for each system and a maintainability survey was conducted on these systems. Their regression analyses were verified against six other software systems not included in the original eight systems which produced the model. The

intention of the work was to determine which software metrics are good predictors of maintainability.

The study presented a one-, four- and five-metric polynomial model (a metric per term) for predicting maintainability. These regression models were assessed for their accuracy in predicting maintainability under a wide range of conditions, aiming to ensure that an excess of one of the factors results in an over- or under-prediction of maintainability. Oman and Hagemeister modified their four metric polynomial in the study by Coleman et al. (1994) to instead use the average Halstead volume over the average Halstead effort, citing "that the volume is a non-decreasing function with concatenation". The paper therefore defines maintainability as:

$$\begin{aligned} MI &= 171 \\ &\quad - 5.2 \times ln(HV) \\ &\quad - 0.23 \times CC \\ &\quad - 16.2 \times ln(LOC) \\ &\quad + (50 \times \sin(\sqrt{2.46 \times COM})) \end{aligned} \qquad (1)$$

The symbols used in Equation 1 are indicated in Table 1.

Table 1: Maintainability index symbols used in Equation 1

Symbol	**Explanation**
HV	Halsteads' volume
CC	Cyclomatic complexity
LOC	Average LOC per module
COM	Average comments per LOC

Halstead Volume

The Halstead volume is not used directly in this study, but is instead used as part of the MI. The Halstead volume is one of a set of metrics introduced by Howard Halstead in 1977. It is the product of the Program Length and the logarithm of the Program Vocabulary, which are functions of the number of operands and operators (Halstead 1977).

In general, operands are variables and constants whereas operators are everything else, and program vocabulary is the sum of the distinct operators and operands.

The Halstead volume will increase linearly with the length of the program and logarithmically with each new concept introduced. The Halstead volume therefore provides a measure of the size of the program independent of language and character set.

Excluding Generated Code

The weighbridge application makes use of an in-house code generator to provide domain classes and basic architectural concerns such as object-relational mapping and service contracts. This means that a significant portion of the create, read, update and delete (CRUD) functions are managed by the generated code for both the server and client side of the application.

Visual Studio does not distinguish between generated and non-generated code when calculating the metrics. In order to prevent the generated code from skewing the results, these should be filtered out. To filter these methods out, the class-level metrics must be recalculated excluding any generated methods.

Cyclomatic complexity and LOC are additive for each class; the MI, however, is not. This is because it is dependent on logarithms of the LOC and the Halstead volume, but is also computed using average values at class, package, and assembly levels, and has a non-additive result based on those measures.

When recalculating the MI at assembly level, excluding generated code, the LOC and cyclomatic complexity are therefore readily available, but the Halstead volume is not. However, the Halstead volume can be recovered from the MI using the cyclomatic complexity and LOC:

$$HV = e^{MI'} \qquad (2)$$

where

$$MI' = \frac{171 - 0.23 \times CC - (16.2 \times ln(LOC)) - 1.71 \times MI}{5.2}$$

Although Visual Studio uses a MAX function to clamp the MI, in practice this has no effect unless the MI is less than zero, which is very unusual and does not occur in the data set for the current case study.

The formula in Equation 2 can be verified by using the calculated Halstead volume to recover the identical MI at both base and aggregated levels. The aggregated MI can therefore be recalculated excluding specific type members from the assembly.

In order to determine which class methods to exclude from the metric data, each method in the assembly indicates the source file from which it is compiled. Since all generated code lives under a generated subdirectory in each project, generated code can be

excluded by means of filtering all methods whose source file contains “generated” in the file path.

Application-Specific Metrics

Aspects of the configuration files used by the weighbridge application have also been measured. These measurements count

7 the number of rules that are defined,

8 the number of core and plug-in extensions that are defined, and

9 the number of LOC in the plug-in assemblies.

Functional Quality Metrics

Functional quality is determined by the absence of defects and can therefore be considered a measurement of the presence of functional quality.

The project team used Jira to assist with planning and tracking the work performed for development and bug fixes. When a bug is logged in Jira, a note is made against a corresponding customer for whom the issue is relevant. The following data is extracted from the Jira (Atlassian 2018) database:

- the type of issue (task, bug); and
- the specific customer for whom this issue is relevant.

The data collected spans approximately 16 months of operational data for the product. The functional quality metric is defined as the count of bugs allocated to each customer or to the core product.

Results

Analysis by Customer

The number of bugs logged per customer was sourced from the Jira data. The number of bugs per customer is shown in Table 2. The other metrics gathered are summarised by customer in Table 3 and all of these metrics are illustrated in Figure 2.

Table 2: Bug counts per customer

Customer	**Bug Count**
Out of the box	38
Heavy customisation	123
Domain modified	25

Table 3: Counts collected for rules per customer

Customer	Plug-ins	Rules	LOC	View-Control Settings	View-Model Settings	Extensions
Out of the box	2	385	1 572	164	253	16
Heavy customisation	28	642	205	492	516	115
Domain modified	31	362	3 587	198	522	44

The LOC refers to the LOC for plug-in assemblies.

The LOC in Table 3 for C2 is seemingly low since the rules defined refer mostly to code that is part of the core application as opposed to code in a custom assembly. Specifically, although the rules defined are specific to the customer, the code implementing those rules is not.

In an analysis of the three types of customer, a strong correlation between the number of rules and bugs is present whereas the MI seems to be relatively unaffected by the number of rules added, as illustrated in Figure 3.

Given the data, adding more rules appears to cause more bugs in the system. However, correlating the number of bugs to the quantity of code in the custom configurations presents with an inverse correlation. This initially seems surprising since the hypothesis is that more custom code leads to more bugs owing to the possibility of introducing errors.

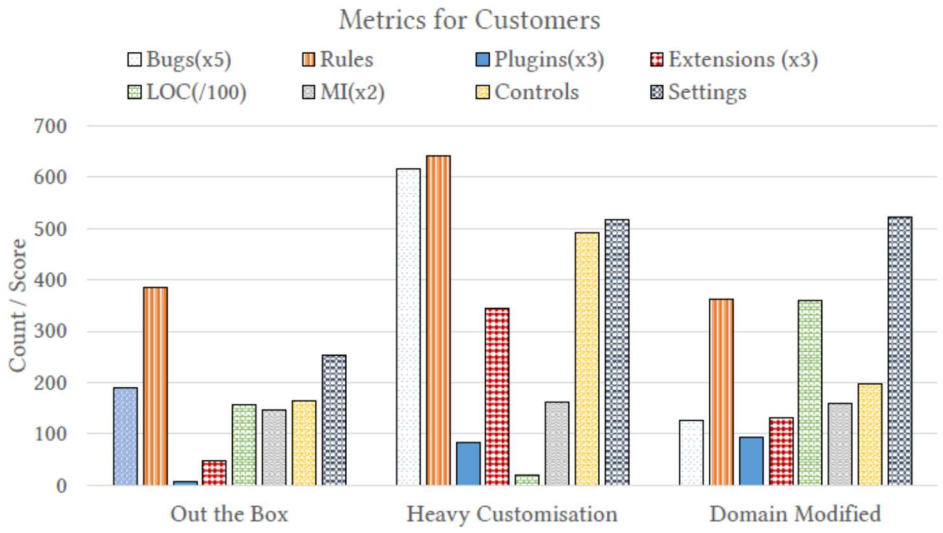


Figure 2: Summary of metrics for each customer
Note: The multipliers are provided in order that the metrics can be plotted on the same range.

In order to understand the observed results, it is noted that more custom plug-ins result in fewer bugs. Inspecting the files which are modified to resolve the bugs, indicates that the files updated are in the core and not the plug-ins. One can conclude that writing code to specifically accommodate a rule for a particular customer results in fewer bugs than would otherwise be introduced if core were made configurable.

Another interesting result is that C3 has the lowest number of bugs, although it was expected that C3 would have the highest number of bugs. C1 has fewer bugs than C2 as expected. The expectations were driven by the degree of changes made for each customer, and C3's changes include a change to the system domain. The data also presents a very high correlation between the number of bugs and number of commits. This is expected because these commits were allocated against the customer using bug issue number. By virtue of this methodology, bugs cause commits.

It was expected that the negative effects on quality would have an exponential component, although the results plotted in Figure 3 indicate an apparent linear relationship with the number of rules changed. This suggests that there is a low dependency between rules in the system. This is validated by noting the system's decoupled module design. Individual extensions are specific to modules and are not shared between modules.

Figure 3: The correlation between bugs, the maintainability index, and line of plug-in code with rules

Spearman's Rank Correlations

A Spearman's rank correlation can also be used to correlate the metrics across customers. This approach is appropriate as the distributions of the measures are non-normal distributions. This approach was taken by Bijlsma et al. (2012) when dealing with defects. It is noted that the Spearman's rank correlation correlates increasing and decreasing rank trends and not the metrics themselves. Given that there are only three samples (and the ranks are therefore 1, 2 and 3), it is expected that correlations will be rounded to factors of 50 per cent. (See Table 4.) The most significant correlation results are, as before:

- An increase in rules strongly correlates with an increase in bugs.
- An increase in the amount of customer-specific code has a strong correlation with a decrease in the number of bugs – suggesting that customer-specific code is more successful at correctly accommodating customer business needs.
- An increase in the amount of customer-specific code also correlates strongly with a decrease in the number of rules – suggesting that changes are accommodated using either code or rules.

Table 4: Spearman's rank correlations for gathered metrics

	Bugs %	**Rules %**	**Plug-ins %**	**MI %**	**LOC %**	**View-Control Settings %**	**View-Model Settings %**
Rules	100	–	–	–	–	–	–
Plug-ins	−50	−50	–	–	–	–	–
MI	50	50	50	–	–	–	–
LOC	−100	−100	50	−50	–	–	–
View-control settings	50	50	50	100	−50	–	–
View-model settings	−50	−50	100	50	50	50	–
Extensions	50	50	50	100	−50	100	50

Plug-In Code Compared to Core Code

This section presents a further analysis that attempts to glean insight regarding the structural quality of the customisations as compared to the core code. The quality of code is compared in the following three areas:

- core – the application code common to all customers;
- plug-in – the customer-specific plug-in code; and
- other – third-party assemblies and non-domain-specific infrastructural or utility code in the product.

Size Comparison

A comparison of sizes for the different areas is indicated in Table 5. When comparing the amount of core and plug-in code it is evident that the plug-in code constitutes about a third of the total. This is substantial but unsurprising given that considerable effort has been put into making the application configurable.

Table 5: Lines of code by category for the weighbridge application

Area	**LOC**	**Assembly Count**
Core	11,790	32
Plug-in	5,364	6
Other	60,586	43

Quality Comparisons

Considering the code metrics discussed, it is noted that the MI is relatively constant across all three areas, although slightly lower for plug-ins as shown in Table 6.

Table 6: The maintainability index averaged over the customers under study

Area	**MI**
Core	90
Plug-in	81
Other	91

Lanza and Marinescu (2007) provide ranges for a number of code metrics which are used to determine whether the metric is rated as high, medium or low based on their analysis of an array of different software projects. Table 7 provides a count of the number of assemblies at each rating. For example, four of the plug-in assemblies are rated as having a high cyclomatic complexity, one is of medium complexity, and one is of low complexity.

With regard to cyclomatic complexity, all code other than plug-in code is rated as high. This suggests that the code base is complex and difficult to understand. On the other hand, only 66 per cent of plug-in assemblies are rated as having a high complexity.

The plug-in code includes modules that have longer methods with higher levels of fanout per call, and more calls per method when compared with other code.

Table 7: Metric ratings per assembly for all selected customers

Area	High	Medium	Low
Cyclomatic complexity per LOC			
Core	32	0	0
Plug-in	4	1	1
Other	43	0	0
LOC per method			
Core	0	0	32
Plug-in	1	1	4
Other	0	0	43
Calls per method			
Core	8	9	15
Plug-in	5	0	2
Other	2	4	37
Fanout per call			
Core	2	8	22
Plug-in	2	0	4
Other	16	1	26

Conclusion

The structural quality effects are measured using the MI and the structural quality remains relatively unaffected by the amount of customisation across the three customers.

It is noted, however, that the MI of plug-in code is at a marginally lower level than the rest of the system as can be seen in Table 6. Given that the complexity of the plug-in code is lower than the core code (see Table 7), it can be inferred that the decreased MI is caused by a combination of longer methods and higher fanout and calls. These would result in an increased Halstead volume and reduce maintainability.

With regard to the functional quality effects, the results indicate that defect counts are lower in cases where fewer rules are defined and more customer-specific code is implemented than for cases where more rules are defined and there is less customer-specific code.

It is relevant that the plug-in code is typically of lower complexity than core code (see Table 7). This may explain why a larger amount of customer-specific code correlates strongly with fewer bugs raised.

An argument can be made that the increased complexity of the core code is a product of the attempt to accommodate many business requirements through configuration. The extensibility in the core code has a negative impact on the functional quality of the application. This creates opportunities for functional errors from unforeseen side effects owing to conflicting customisation choices. In summary, the effects on quality are mixed and the architecture can only be regarded as being partially successful in supporting customer changes.

This work would benefit from a broader investigation involving additional customers. It could be extended to include open-source projects which have a high degree of configurability and healthy issue tracking.

References

Apache Software Foundation. 2018. "Apache™ Subversion®." https://subversion.apache.org/.

Atlassian Pty Ltd (ABN 53 102 443 916). 2018. "Jira/Issue and Project Tracking Software." https://www.atlassian.com/software/jira.

Basili, Victor R., and H. Dieter Rombach. 1988. "The TAME Project: Towards Improvement-Oriented Software Environments." *Software Engineering*, IEEE Transactions on 14 (6): 758–73. https://doi.org/10.1109/32.6156.

Bijlsma, Dennis, Miguel Alexandre Ferreira, Bart Luijten, and Joost Visser. 2012. "Faster Issue Resolution with Higher Technical Quality of Software." *Software Quality Journal* 20 (2): 265–85. https://doi.org/10.1007/s11219-011-9140-0.

Chappell, David. 2011. "The Three Aspects of Software Quality." Accessed 17 September 2017. http://davidchappell.com/writing/white_papers/The_Three_Aspects_of_Software_Quality_v1.0-Chappell.pdf.

Coleman, Don, Dan Ash, Bruce Lowther, and Paul Oman. 1994. "Using Metrics to Evaluate Software System Maintainability." *Computer* 27 (8): 44–49. https://doi.org/10.1109/2.303623.

Easterbrook, Steve, Janice Singer, Margaret-Anne Storey, and Daniela Damian. 2008. "Selecting Empirical Methods for Software Engineering Research." *Guide to Advanced Empirical Software Engineering* 285–311. https://doi.org/10.1007/978-1-84800-044-5_11.

Fenton, Norman E., and Shari L. Pfleeger. 1998. *Software Metrics: A Rigorous and Practical Approach*. Brooks. Cole.

Fowler, Martin. 2004. "Inversion of Control Containers and the Dependency Injection Pattern." https://martinfowler.com/articles/injection.html#ServiceLocatorVsDependencyInjection.

Freimut, Bernd, Christian Denger, and Markus Ketterer. 2005. "An Industrial Case Study of Implementing and Validating Defect Classification for Process Improvement and Quality Management." In *Software Metrics, 2005. 11th IEEE International Symposium*. IEEE, 10–19. https://doi.org/10.1109/METRICS.2005.10.

Guido, Capaldo, and Rippa Pierluigi. 2008. “A Methodological Proposal to Assess the Feasibility of ERP Systems Implementation Strategies.” In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, 401–401. https://doi.org/10.1109/HICSS.2008.30.

Hall, Tracy, and Norman Fenton. 1997. “Implementing Effective Software Metrics Programs.” *IEEE Software* 2: 55–65. https://doi.org/10.1109/52.582975.

Halstead, Maurice H. 1977. *Elements of Software Science. Vol. 7*. New York: Elsevier.

Kelemen, Zádor D., Gábor Bényász, and Zoltán Badinka. 2014. “A Measurement Based Software Quality Framework.” Preprint arXiv:1408.3253v2.

Kresowaty, Jason. 2008. *FxCop and Code Analysis: Writing Your Own Custom Rules*.

Kumar, M. N. Vijaya, A. V. Suresh, and P. Prashanth. 2009. “Analyzing the Quality Issues in ERP Implementation: A Case Study.” In *Emerging Trends in Engineering and Technology (ICETET), 2009 2nd International Conference on*. IEEE, 759–64. https://doi.org/10.1109/ICETET.2009.34.

Lanza, Michele, and Radu Marinescu. 2007. *Object-Oriented Metrics in Practice: Using Software Metrics to Characterize, Evaluate, and Improve the Design of Object-Oriented Systems*. Springer.

Light, Ben. 2001. “The Maintenance Implications of the Customization of ERP Software.” *Journal of Software Maintenance and Evolution: Research and Practice* 13 (6): 415–29. https://doi.org/10.1002/smr.240.

Microsoft Contributors. 2018. “FxCop – What’s New in Visual Studio 2013.” https://blogs.msdn.microsoft.com/devops/2013/07/03/what-is-new-in- code-analysis-for-visual-studio-2013/. Version 12.021005.1.

Microsoft Corporation. 2018. “Visual Studio 2013.” https://visualstudio.microsoft.com/. Version 12.

Oman, Paul, and Jack Hagemeister. 1994. “Construction and Testing of Polynomials Predicting Software Maintainability.” *Journal of Systems and Software* 24 (3): 251–66. https://doi.org/10.1016/0164-1212(94)90067-1.

Razina, Ekaterina, and David S. Janzen. 2007. “Effects of Dependency Injection on Maintainability.” In *Proceedings of the 11th IASTED International Conference on Software Engineering and Applications*, Cambridge, MA, 7.

Smith, Josh. 2009, February. “The Model-View-ViewModel (MVVM) Design Pattern for WPF.” *MSDN Magazine*.

Van Deursen, Arie. 2014. “Think Twice before Using the ‘Maintainability Index’.” Accessed 17 September 2017. https://avandeursen.com/2014/08/29/think-twice-before-using-the-maintainability-index/.

Supporting Trainee Teachers of Computer Science with Game Authoring Tools

Jecton Tocho Anyango
https://orcid.org/0000-0002-7295-2137
School of Information Technology,
University of Cape Town, South Africa
ANYJEC001@myuct.ac.za

Hussein Suleman
https://orcid.org/0000-0002-4196
School of Information Technology,
University of Cape Town, South Africa
hussein@cs.uct.ac.za

Abstract

Despite the current evidence suggesting the potential of game-based learning in education, developing serious games remains difficult, time-consuming and expensive. This leads to low adoption of game-based learning in mainstream teaching. In particular, trainee teachers of Computer Science are not likely to develop and adopt serious games when they just begin teaching programming. To deal with this problem, we designed and conducted empirical evaluations of a prototype game authoring tool called the Recursive Game Generator. The tool is aimed at supporting these teachers who have limited game-programming skills. A total of 22 trainee teachers of Computer Science evaluated the Recursive Game Generator using the standard AttrakDiff questionnaire. A good user experience was indicated through the results of the scale mean scores. The mean values of word pairs showed that the participants found the prototype (a) usable for achieving goals, (b) supporting novelty, content and stimulation, and (c) attractive. In addition, 70 per cent of the participants found the approach of game generation a good idea for aiding teachers of Computer Science, whereas 65 per cent noted that the generated games were interactive, practical, interesting and fun, demonstrating the tool's potential educational value. Consequently, the findings from this study may provide an opportunity for inexperienced computing teachers to embrace the idea of game generation to support the teaching and learning of difficult concepts in the first year of Computer Science.

CCS concepts: social and professional topics, computer science education, CS1
Keywords: AttrakDiff, CS, CS1, trainee teachers, game design, game generator, authoring, UX, user experience, recursion

Introduction

Despite existing evidence suggesting the potential of game-based learning (GBL) in education (Papastergiou 2009), developing serious games remains difficult, time-consuming and expensive (Green et al. 2018; Torrente et al. 2008). Consequently, there

is a low adoption rate of GBL in mainstream teaching (Tang and Hanneghan 2010). Trainee teachers (trainees), in particular, are not likely to develop games and to adopt serious games when they embark on teaching programming. To deal with this problem, some researchers have proposed game authoring tools (Khenissi, Essalmi, and Jemni 2015; Marchiori et al. 2012; Osborn et al. 2019; Pérez-Colado et al. 2019; Torrente et al. 2010). However, very limited studies have tested the idea of game authoring in Computer Science Education (CSE) – particularly with trainees. In this paper, we attempt to narrow this gap. We test the idea of a game generator tool to support CS trainees who may need such an intervention more during their teaching practice or immediately when they start teaching. We designed a prototype of such a tool called the Recursive Game Generator (RGG) (https://programmingwithfun.net/) and elicited useful feedback from prospective users (trainees) concerning their user experience (UX), which could lead to adopting GBL. Furthermore, we assessed the participants' overall subjective opinions about the prototype and the potential efficacy of such a programming teaching tool that is based on the concept of a game generator to support trainees.

The aim was to test how people who may need such a tool more to aid their teaching practices will react to it. In the case study, we used the recursion topic given its difficulty among learners (Malik and Coldwell-Neilson 2017; Miller, Settle, and Lalor 2015; Scholtz and Sanders 2010). The participants used a prototype game generator tool to create and play custom games from two given game examples. This is a second user study. The first used experienced CS teachers (Anyango and Suleman 2021) whereas the current used trainees. The prototype could support both high school and higher education (tertiary) teachers. In this paper, we deal with the following two questions:

- What is the adoption potential of an innovative support tool that uses the concept of a game generator to help CS trainee teachers to create games that can teach programming?
- What is the overall opinion of CS trainee teachers about the RGG and the generated games?

The contribution of the current work is threefold:

- it provides empirical evidence that the prototype game generator tool could be used by CS trainee teachers, which could lead to the potential of adopting GBL;
- it provides preliminary insights into the potential educational value of the generated games; and
- it highlights the RGG features' design gaps and ideas for improvements that could lead to design implications for other tools for teaching programming.

Related Work

Difficulty of Teaching Programming

Teaching introductory programming to novice students is difficult (Aedo Lopez et al. 2016; Weir et al 2005), especially regarding concepts such as abstraction, functions, recursion and reuse (Aedo Lopez et al. 2016; Eagle and Barnes 2009). A previous study found that trainee teachers, in particular, are more likely to find it difficult to teach programming (Major, Kyriacou, and Brereton 2011). When asked how difficult it is to teach programming concepts to high school students, 60 per cent of the CS trainee teachers said that it is difficult whereas less than 50 per cent noted that they had the confidence to teach the subject (Major, Kyriacou, and Brereton 2011).

Supporting CS Trainee Teachers

Although GBL has been proposed as an alternative innovative approach to teach programming in higher education, developing educational games is difficult, time-consuming and expensive (Green et al. 2018). To deal with this problem, some works have fronted game authoring tools to support non-technical domain experts (teachers) (Osborn et al. 2019; Pérez-Colado et al. 2019). However, the absence of appropriate authoring environments and support for teachers inhibit many who wish to adopt GBL approaches in mainstream teaching (Tang and Hanneghan 2010). Some previous work found that CS teacher interest in a new approach could drive the adoption (Ni 2009). Supporting this claim, another study has also suggested that trainee teachers are more likely to adopt teaching methods and philosophies based on their own experiences (Ladd and Harcourt 2005). The current study builds on a previous work that attempted to understand the potential of a programming teaching tool to support prospective high school teachers (trainees) (Major, Kyriacou, and Brereton 2011), but from a different perspective. Whereas the former work investigated the effectiveness and potential of a robot teaching tool that simulates programming, the current study evaluates the potential of a prototype tool for game authoring to support CS trainees.

Designing Serious Games

Klemke et al. (2015) reported the absence of standards for serious games design. Regarding game research, Kultima (2015) observes that most studies on games do not emphasise the notion of design research. Regarding pedagogy, Walliaka et al. (as quoted in Malliariakis and Satratzemi 2014) suggest that existing educational games focusing on computer programming do not enable teachers to configure the game environment according to the pedagogical goals of the respective unit of learning. Supporting this claim, Medeiros, Ramalho and Falcão (2019) also report the lack of scaling and personalised teaching as some of the challenges CS1 teachers face. Inherently, these pose another threat to the adoption of GBL as most educators always prefer being in control of their learning material by creating, modifying, reusing and sharing content (Marchiori et al. 2012).

Evaluating Game Authoring Tools

Tornero et al. (2010) developed e-Training DS – an authoring tool for integrating portable games for computer science in e-learning. The tool allows instructors to maintain a library of mini games that can be easily created by customising generic game patterns. In a case study, the time taken to create and modify the game and its assessment was measured. The results showed that the process was feasible for an educator.

Marchiori et al. (2012) evaluated the system of Writing Environment for Educational Video games (WEEV) – a game authoring platform built on <e-Adventure>. Three evaluations were conducted, namely, the formative evaluation, end-user evaluation, and game creation. The first two evaluations did not involve creating educational games. Instead, they assessed the impressions and perceptions of the educators. The results from the formative evaluation with 20 students found some information in the system excessive and the software not useful. The end-user evaluation found that the tool improved the game creation process of nine educators but noted that the tool was complex to use.

Pérez-Colado et al. (2019) conducted the first preliminary evaluation of users' experiences of the uAdventure authoring platform. The tool was evaluated by heterogeneous users: (i) non-technical users (teachers) ($n = 2$); (ii) artists ($n = 2$); and (iii) programmers ($n = 6$) with different degrees of technical knowledge. The aim of the evaluation was to identify issues or difficulties with the tool. The difficulty of each performed task was measured on a scale of four levels: easy, normal, difficult and not accomplished (Pérez-Colado et al. 2019). The results showed that most tasks (65%) were rated as simple. Nonetheless, the participants identified 29 issues or difficulties with the tool for further design considerations.

The reviewed works suggest that limited support has been given to CS teachers with game authoring tools. They also highlight educational games design gaps – particularly those created from game authoring platforms. Finally, they reveal that limited empirical evaluations have been conducted to test the idea of game generation with the potential beneficiaries such as CS trainee teachers. The current work attempts to narrow these gaps.

Prototype Design

Design Theories

Hidden Complexity Theory

The complexity theory hides the complexities of the underlying technologies from the users (in this case the teachers). This theory supports different levels of complexity in that novice users of a tool can easily design almost any instance of an artefact by merely selecting from given examples and customising various parameters. At the same time, experienced users can access advanced system features and create complex artefacts.

On the issue of the complexity of authoring tools, Murray (2016, 2) poses two questions: (i) "Who are going to use these tools?" and (ii) "How do we make sure that the tools meet end-user needs?" In our case, if the intention is to support advanced users who may wish to add advanced pedagogical content, then the tool's usability would be compromised, particularly, in regard to novice users (Mirel 2004). On the other hand, if design purely targets novice users who are only interested in serious games with simple pedagogical content, then complexity may suffer at the expense of usability (Murray 2004). If we are to design for both types of user, a trade-off is required (Mirel 2004; Murray 2004, 2016).

In another work, Murray (2016, 3) raises the issue of "how one matches the complexity of the authoring task to the complexity of a tool and the complexity-capacity of the target user". Complexity–capacity or cognitive complexity of a user is a "person's capacity to perform complex mental or behavioral tasks" (Murray 2016, 18). He goes on to propose four theoretical foundations that could inform future authoring tools design, namely, complexity in software design, the activity theory, epistemic forms and games, and the theory on adult cognitive development (Murray 2016). The hidden complexity theory was chosen to guide the design of the authoring tool prototype because it supports the design of simple tools targeting novices and also gives the flexibility to allow experienced users to explore complex system features (Murray 2016; Soloway, Guzdial, and Hay 1994).

Differentiating Interfaces Theory

A study conducted by Karoui, Marfisi-Schottman and George (2017) noted that teachers found authoring tools for mobile learning games either too poor to create games that could fit their teaching needs or too complex to use. To solve this problem, the authors proposed a design model for authoring tools that supports several conceptual levels. The rationale is to design different interfaces that target different user profiles (novice and experienced tool users). A nested-design approach comprising novice, intermediate, and advanced modes is suggested. With the novice mode, novice programming teachers are able to easily author simple learning games to experiment with students in class. To gain further experience with the authoring tool, the design approach gradually accords such users more features at the intermediate mode. Lastly, the advanced mode allows more experienced teachers to explore advanced features when creating more complex game instances. The differentiating interfaces theory ensures the design of different interfaces or user profiles for different users (novices and advanced users) Karoui, Marfisi-Schottman and George (2017). This theory is considered relevant since the proposed tool is envisaged to support both experienced teachers (Anyango and Suleman 2021) and the inexperienced such as the trainee teachers in the current study.

Design Methodology

Given the absence of standards for serious games design, the design process of the RGG adopts the proposals by the working group for the Game Development for Computer

Science Education at the annual conference on Innovation and Technology in CSE (Johnson et al. 2016) and those by Saavedra et al. (2014). More detail is reported in the study by Anyango and Suleman (2021). When it comes to pedagogy, the design approach ensures that custom games created from the given game examples have learning tasks aligned with goals.

For instance, Table 1 illustrates aligning one of the games (the DnD game) levels with pedagogy in CS1. The design focuses on the teaching and learning of the difficult recursion topic through common examples or scenarios used by most instructors. The students are given the learning tasks which they solve in a game environment. This makes the whole experience fun, challenging, engaging and motivating (Wicentowski and Newhall 2005). The students learn by writing programs that implement a critical aspect of the game or act as a player in an existing game (Hakulinen 2011). The design also affords teachers the ability to scale and personalise their teaching. Educators have the flexibility to easily configure the game environment elements such as view, scene, background image, and wall sprites during the generation process.

Table 1: Alignment of the DnD game levels with pedagogy in CS1

Level	**Learning outcome/task**	**Programming concept**	**Python code snippet**
1	Complete the given code so that the function returns double the given parameter x	Function	def double(x) : #insert code here
2	Complete the given code so that the swingsword function returns the factorial of 0 (0) if the parameter n is smaller than 1 otherwise it returns 0	Recursion – parameters, base case, recursive call	def swingsword(n) : #insert code here

3	Complete the function swingsword so that it calls itself but only if n > 1, otherwise it should return the factorial of 0 (0)	Recursive call with stop condition	x=0
			def swingsword(n) :
			global x
			if(n<1) :
			  return 1
			else
			  x += 1
			  n −= 1
			#insert recursive call here

4	Complete the code snippet so that the function swingsword returns the factorial of the parameter n given to it (n)	Recursive call	recursiveCalls=0
			def swingsword(n) :
			  global recursiveCalls
			  recursiveCalls += 1
			#insert code here

The RGG is a web-based tool (https://programmingwithfun.net/) with options that produces a customised download. The application logic contains HTML, Java Script, PHP, and C#. The Unity Game engine is used. C# scripts send and access data between the application and the back-end server. MySQL database is implemented in the back end. In the front end, the user (instructor) sees and interacts with the file system through a web browser. The PHP code in the back end validates and verifies user requests. The custom games created from the Mushroom Picker game example enable novice students to visualise execution of program codes as presented in Chaffin et al. (2009) and Cooper, Dann and Pausch (2000). Furthermore, they allow students to use simple commands such as UP(), DOWN(), LEFT(), RIGHT() (Aedo Lopez et al. 2016; Dann and Pausch 2000) which make programming easy for novices. More generated game-design aspects are reported in the study by Anyango and Suleman (2021).

Evaluation

Empirical Study

The prototype game generator (the RGG) was empirically evaluated in a controlled laboratory user study. A within-subject design was employed. The participants were asked to evaluate their UX when creating custom games using the RGG. The games are created from two game examples given to users (the DnD game and the Mushroom Picker game). In the study we investigate the subjective usability and user experiences (dependent variables) of the participants.

Participants

The participants of the main study were 22 CS trainee teachers from the faculty of education. The subjects are training to be CS teachers in high school. Therefore, one of their teaching subjects is computing. The participants were drawn from Kenyatta University (a Kenyan University known for training high school teachers). The target participants were motivated by the fact that (i) at the time of the experiment, most universities in the country have resumed face-to-face learning after eight months of closure owing to the Covid-19 pandemic, and (ii) the ethical clearance process is faster in Kenya than in South Africa. In addition, it was assumed that the target participants (i) have been taught some programming courses, (ii) have learnt how to plan teaching computing in high school, and (iii) would definitely encounter the challenges of teaching programming (especially the difficult topic of recursion).

The sample varied by age, teaching subjects, interest in games, and learning with game experience, among other things. A total of 63 per cent of the participants fell in the 18 to 20 age bracket. A total of 85 per cent had registered for mathematics and computing as their 2 teaching subjects. A total of 77 per cent considered themselves gamers. A total of 95 per cent had not been taught any course using games before and 60 per cent did not study computing in high school.

Materials

The following materials were used in the study:

- an online informed consent form;
- a printout of the task sheet;
- a video tutorial;
- a computer with a web browser and internet connection;
- the AttrakDiff 2 questionnaire; and
- an exit survey.

Research Team

The research team comprised two people, namely, the first author and a volunteer research assistant. The research assistant was a CS high school teacher from Kenya who had interacted with the prototype in a prior user study.

Tasks

Tasks performed by the participants included:

- log in using a given user account and password;
- create custom static or dynamic game by customising assets;
- download the generated game;
- customise the generated game;
- play test the generated game;
- complete the AttrakDiff 2 questionnaire; and
- complete the exit survey.

Measurement Items

The game generation and play experiences were measured through subjective ratings of the (i) usability, (ii) user experience, and (iii) qualitative statements. The summative measures were taken from a suitable sample of potential users who performed given tasks in a realistic context of use (Bevan 2008). The AttrakDiff questionnaire developed by Hassenzahl (2004) was used to assess the perceived pragmatic quality, the hedonic quality, and the attractiveness of the prototype game generator. The questionnaire has four scales with a total of 26 items. The items that are on a scale from −3 to +3 are measured as a semantic differential. Each item consists of a pair of terms with opposite meanings.

Procedure

Ethical Clearance and Study Planning

Ethical clearance was first sought from the University of Cape Town. This was followed by acquiring two more research approvals, namely a research permit from the National Commission for Science, Technology and Innovation (NACOSTI) in Kenya, and an approval letter to conduct research at Kenyatta University in Kenya. The first author established rapport with two contact persons (one lecturer from the Faculty of Education and another from the Department of Computer Science at Kenyatta University).

During every visit to the University, the research team strictly adhered to the protocols laid down to mitigate the spread of the Covid-19. Some of these included (i) registering with the campus security officers manning the main entrance gate, (ii) frequently washing and sanitising hands, (iii) wearing face masks, and (iv) keeping social distancing. At the Department of Computer Science, the research team held two meetings with the contact lecturer and the class representative to plan the study. The lecturer had just concluded practical programming classes with the learners in a CS1 course. The lecturer had taught basic programming concepts such as variables, loops, conditionals, procedures, functions and lists. The concept of recursion had also been introduced.

A follow-up meeting was later scheduled with the target participants in the laboratory. During the meeting, the lecturer introduced the research team. Before each meeting, all participants washed their hands using the facility at the entrance of the department's building and sanitised inside the laboratory. The team was neutral to the participants as none of the members was a lecturer at Kenyatta University. The team then explained the purpose of the study and demonstrated the way in which the prototype game generator works, using a video. This was followed by requesting the trainee teachers to participate in a study to evaluate the prototype. Through the class representative, the researcher announced the study in the class mailing list and other social media platforms. Out of 32 registered CS trainee teachers, 27 willingly signed up to participate. A week before the main experiment, two further visits were made to the department to secure and set up the laboratory.

Pilot Study

Next, a pilot study was conducted with 5 out of the 27 volunteer participants. The 5 did not participate in the final main experiment. They used the prototype to create custom games from the DnD game example and completed the AttrakDiff 2 questionnaire. The aim was to ensure that (i) the Lime Survey and Amazon servers (in which the prototype was hosted) were stable, (ii) the participants understood the AttrakDiff 2 questionnaire terminology, and (iii) the questionnaire could be completed in a reasonable amount of time.

Main Experiment

A laboratory fitted with full internet access was setup for the main experiment. The same laboratory used to teach the students during practicals was used. This gave the participants a familiar environment (Lallemand, Gronier, and Koenig 2015). A total of 22 participants who did not participate in the pilot study were divided into two equal groups (11 students each). Each participant was randomly assigned to one of the groups. Before the experiment, the team ensured that each participant wore a face mask and sanitised. A printout of the task sheet was then given to each participant. After that, the research team again explained the purpose of the experiment and the tasks to be performed. Each participant signed an online informed consent form. The first group (Figure 1(a)) started by using the prototype to create custom serious games from the DnD game example, then played the generated games. This was followed by completing an online AttrakDiff 2 questionnaire. Next, the group generated serious custom games from the Mushroom Picker game example. Afterwards, they played the generated games, then again completed an online AttrakDiff 2 questionnaire. Finally, every participant of group 1 completed an online exit survey comprising demographic information and two optional open-ended questions. These questions allowed the users to further express any personal general views about the prototype and any negative experiences for future improvements. During the experiment, other than the participants, only the first author and the research assistant were present. Their roles included clarifying the tasks to the participants and answering queries. Before the participants of group 1 left the laboratory, the first author thanked them for their time.

This procedure was repeated for the second group (Figure 1(b)) until all 22 participants completed evaluating the prototype. However, for the second group, the order was reversed. The participants first generated custom games from the Mushroom Picker game example and then from the DnD game. The two groups did not interact with each other during or immediately after the first experimental laboratory session. Each test session lasted for approximately 1 hour and 40 minutes.

(a) Group 1 participants

(b) Group 2 participants

Figure 1: Experimental laboratory sessions

Data Collection and Analysis

Both quantitative and qualitative UX data were collected through user self-reporting. Three surveys were answered by each participant. The first and the second used Likert scales to collect overall subjective user experiences using the standard AttrakDiff questionnaire. The third was an exit survey containing demographic information and two open-ended questions. The first question was about the most negative issues and the second captured overall comments about the prototype and the game generation idea. We used the Statistical Package for the Social Sciences (SPSS) to compute and analyse the mean score for each scale of the AttrakDiff. Mean values of the word pairs were computed and analysed. The mean values were compared across the pragmatic quality, the hedonic quality, and the attractiveness scales. The evaluation was pegged on a 7-point semantic differential scale (i.e. −3: complicated, +3: simple). For analysis, we gave answers the values: −3, −2, −1, 0, 1, 2 or 3. Text responses from the final user comments were categorised, coded and analysed thematically (Wilson 2019). In addition, the NVivo 12 software was used to analyse the qualitative data.

Findings and Analysis

Scale Mean Scores

We arrived at the mean values for the four AttrakDiff dimensions (scales) by averaging the values of all answers inside each dimension. Table 2 presents the details and the graph in Figure 2 is a visualisation. In the graph, the vertical axis shows the average assessment values of word pairs inside each group whereas the horizontal axis depicts the four word groups or dimensions. The dimensions are pragmatic quality (PQ), hedonic quality identity (HQ-I), hedonic quality stimulation (HQ-S), and attractiveness (ATT).

Overall, the UX reported by the participants suggest a good experience when creating custom games from both the DnD and Mushroom Picker game examples. The graph in Figure 2 shows that the values were above zero for all dimensions. The attractiveness dimension was rated the highest by all the participants for the prototype game generator. This suggests that the participants found the prototype's design attractive and likeable. Moreover, the participants appear to have a positive general impression of the prototype. The PQ results suggest that the participants found it easy to understand how to use the prototype and felt in control of their interactions with it, especially when creating custom games from the dynamic game example.

Table 2: Mean values

Dimension	Static game	Dynamic game
PQ	0.95	1.64
HQ-I	1.36	1.68
HQ-S	1.32	1.68
ATT	1.73	2.01

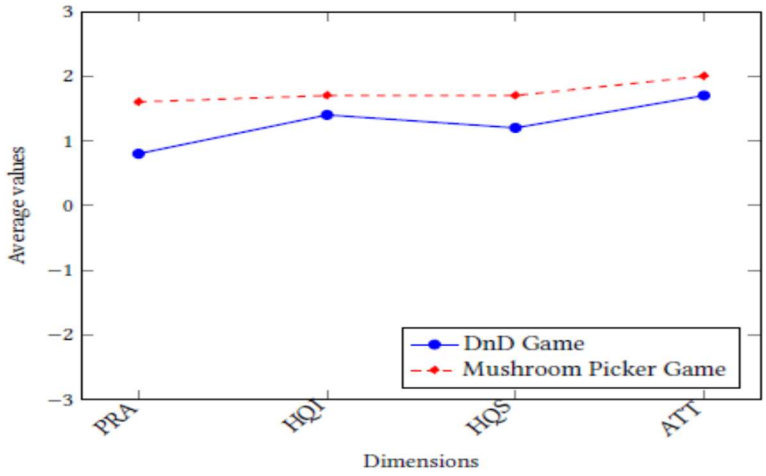


Figure 2: AttrakDiff dimensions average values

The HQ-I and HQ-S mean scores were similar for the Mushroom Picker game example and almost the same for the DnD game. This reveals that the participants found the prototype equally stimulating regarding novelty, content and interaction when using it to create custom games from both game examples. In addition, it suggests that playing the generated games was straightforward, brought players closer to the game play and connected them to other players, particularly for the dynamic Mushroom Picker game example. Results for the HQ-I are promising, given the prominence the CS research community has accorded it (Diefenbach, Kolb, and Hassenzahl 2014).

Mean Values of Word Pairs

Pragmatic Quality

Figure 3(a) presents the mean values of word pairs of PQ and HQ-I dimensions of the AttrakDiff questionnaire. PQ represents the word pairs technical – human, complicated – simple, impractical – practical, cumbersome – straightforward, unpredictable – predictable, confusing – structured, and unruly – manageable. It measures the usability of a product with regard to how successfully a user can use a product to achieve their goals (Marti and Iacono 2015). The findings suggest that using the RGG to create custom games from both the DnD and the Mushroom Picker game

examples gave the participants a positive usability experience. However, the experience was more positive for the Mushroom Picker game example. As can be seen from Figure 3(a), the prototype was clearly structured and manageable. This suggests that the participants were in control of their interaction with it and found the user interface organised.

Hedonic Quality Identity

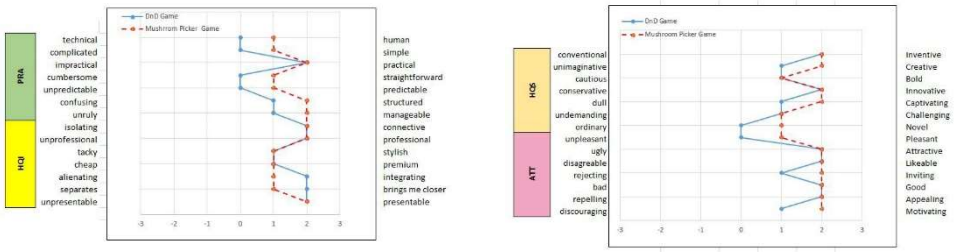
Regarding the HQ-I dimension, the participants rated all the items positively. This indicates that the participants socially identified (Marti and Iacono 2015) with the prototype. This finding is particularly useful given that most participants were in the 18 to 20 and 20 to 22 years age brackets. In addition, they found the prototype connective, professional, stylish and premium. The high perception of HQ-I towards the use of the RGG highlights an interesting finding: young and inexperienced CS teachers found the prototype socially engaging. This could potentially increase acceptance of the game generation idea (Novak and Schmidt 2009) among this group of users who may need it most given their limited teaching experience.

Hedonic Quality Stimulation

The HQ-S indicates the extent to which a product can support user needs relating to novelty, content, stimulation, and presentation of style (Marti and Iacono 2015). Figure 3(b) illustrates that users found the prototype inventive, creative, bold, innovative, captivating and challenging. This finding suggests that the prototype gave users the flexibility of use and the ability to explore alternative ways of effective use.

Attractiveness

Concerning the attractiveness dimension, the participants found the prototype attractive, likeable, appealing and good when creating and playing custom games from both the DnD and Mushroom Picker game examples (Figure 3(b)). The findings reveal a positive overall value of the game generator prototype on the basis of its perceived quality (Marti and Iacono 2015).



(a) Pragmatic and hedonic quality identity (b) Hedonic quality stimulation and attractiveness

Figure 3: AttrakDiff mean values of word pairs

Qualitative Comments

To gather more insights about the prototype, we asked the participants to answer two open-ended questions. The first was about the bugs or technical issues they encountered during the usability test and any design suggestions. The second was on their final opinions. A word-cloud analysis show that words such as game(s), generator, good idea, teaching, teach, programming, help, learners, teacher, and students had the highest percentage. Regarding bugs or technical issues, two themes arose: (i) prototype hosting – internet speed; and (ii) syntax errors. A total of 55 per cent of the participants reported slow download of the generated games when testing the prototype. However, this was not owing to the hosting of the prototypes but rather to the slow internet speed in the laboratory. In addition, some participants had challenges with the Python syntax since they had not encountered programming with the language. One participant suggested the use of different colour code segments to ease code readability and debugging.

Overall, the thematic content analysis revealed that 70 per cent of the participants found the game generator tool a good idea for supporting trainee CS teachers who may wish to adopt GBL during their teaching practice or when they start teaching. Another 65 per cent noted that the generated games were interactive, practical, interesting and fun for learning programming. Some direct quotes follow:

> The generator prototype is a convenient tool for generating games, especially for teachers.
>
> . . . can be used by teachers in teaching because first it is more fun . . .
>
> . . . the game looks interesting and fun to play and easy to generate.
>
> I recommend the game.

Discussion

The positive findings based on experiences of the participants from this study are promising and a clear indication of the potential adoption of the proposed game generation approach in teaching programming in higher education. Similar to findings from a previous study (Ni 2009), the finding that the participants were excited about the game generation idea and the created games suggests that trainee teachers are most likely to adopt the proposed teaching tool.

Regarding development, two issues emerge from the findings that could inform future design of serious games or other programming tools. They include the need to consider (i) different colours in the code snippets design, and (ii) other programming languages. Different colours are perceived to improve code readability and debugging. On the other hand, the finding that the participants had a better UX with the Mushroom Picker game compared to the DnD game confirms earlier work by Chaffin et al. (2009). According

to the authors, students find it fun to learn programming through code visualisation and simulation.

Finally, the fact that all generated games have an integrated development environment could be useful for monitoring student problem-solving activities and coding behaviour (Lyulina et al. 2021).

Conclusions and Future Work

In this paper, we presented the findings from user experience evaluations of a prototype game generator tool called the RGG. We evaluated the prototype with 22 CS trainee teachers. The current study could be limited by the number of participants. However, 22 participants (70%) could be considered representative enough given the Covid-19 pandemic. Moreover, the participants had been taught some programming courses and the way in which to prepare computing classes. Consequently, they were regarded as suitable for the study. Evidence gathered establish that the participants had a good UX with the prototype and found the game generator idea to be good. We therefore argue that supporting CS trainee teachers with such game authoring platforms has the potential of advancing the adoption of GBL in CSE education, particularly among this audience. The proposed idea of a programming game authoring tool could benefit both high school and higher education (tertiary) CS teachers. Our next study will evaluate the effectiveness of the generated games for learning programming through the lens of students.

Acknowledgements

This research was financially supported by the Hasso Plattner Institute for Digital Engineering, the National Research Foundation (NRF) of South Africa (grant numbers: 85470 and 88209) and the University of Cape Town. The opinions, findings, conclusions and recommendations expressed in this publication are solely those of the authors and do not reflect those of the NRF.

References

Aedo Lopez, Marco, Elizabeth Vidal Duarte, Eveling Castro Gutierrez, and Alfredo Paz Valderrama. 2016. “Teaching Abstraction, Function and Reuse in the First Class of CS1: A Lightbot Experience.” In *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*, 256–57. https://doi.org/10.1145/2899415.2925505.

Anyango, Jecton Tocho, and Hussein Suleman. 2021. “Supporting CS1 Instructors: Design and Evaluation of a Game Generator.” In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education*, 115–21. https://doi.org/10.1145/3430665.3456306.

Bevan, Nigel. 2008. “Classifying and Selecting UX and Usability Measures.” In *International Workshop on Meaningful Measures: Valid Useful User Experience Measurement*, 11: 13–18.

Chaffin, A., Katelyn Doran, Drew Hicks, and Tiffany Barnes. 2009. “Experimental Evaluation of Teaching Recursion in a Video Game.” In *Proceedings of the 2009 ACM SIGGRAPH Symposium on Video Games*, 79–86. https://doi.org/10.1145/1581073.1581086.

Cooper, Stephen, Wanda Dann, and Randy Pausch. 2000. “Alice: A 3-D Tool for Introductory Programming Concepts.” *Journal of Computing Sciences in Colleges* 15 (5): 107–16.

Diefenbach, Sarah, Nina Kolb, and Marc Hassenzahl. 2014. “The ’Hedonic’ in Human–Computer Interaction: History, Contributions, and Future Research Directions.” In *DIS ’14: Proceedings of the 2014 Conference on Designing Interactive Systems,* 305–14. https://doi.org/10.1145/2598510.2598549.

Eagle, Michael, and Tiffany Barnes. 2009. “Experimental Evaluation of an Educational Game for Improved Learning in Introductory Computing.” *ACM SIGCSE Bulletin* 41 (1): 321–25. https://doi.org/10.1145/1539024.1508980.

Green, Michael Cerny, Ahmed Khalifa, Gabriella A. B. Barros, Andy Nealen, and Julian Togelius. 2018. “Generating Levels that Teach Mechanics.” In *Proceedings of the 13th International Conference on the Foundations of Digital Games*, 55. https://doi.org/10.1145/3235765.3235820.

Hakulinen, Lasse. 2011. “Using Serious Games in Computer Science Education.” In *Proceedings of the 11th Koli Calling International Conference on Computing Education Research*, 83–88. https://doi.org/10.1145/2094131.2094147.

Hassenzahl, Marc. 2004. “The Interplay of Beauty, Goodness, and Usability in Interactive Products.” *Human–Computer Interaction* 19 (4): 319–49. https://doi.org/10.1207/s15327051hci1904_2.

Johnson, Chris, Monica McGill, Durell Bouchard, Michael K. Bradshaw, Víctor A. Bucheli, Laurence D. Merkle, Michael James Scott, Z. Sweedyk, J. Ángel Velázquez-Iturbide, Zhiping Xiao, and Ming Zhang. 2016. “Game Development for Computer Science Education.” In *Proceedings of the 2016 ITiCSE Working Group Reports*, Arequipa, Peru, 23–44. https://doi.org/10.1145/3024906.3024908.

Karoui, Aous, Iza Marfisi-Schottman, and Sébastien George. 2017. “A Nested Design Approach for Mobile Learning Games.” In *Proceedings of the 16th World Conference on Mobile and Contextual Learning*, Larnaca, Cyprus. https://doi.org/10.1145/3136907.3136923.

Khenissi, Mohamed Ali, Fathi Essalmi, and Mohamed Jemni. 2015. “Comparison between Serious Games and Learning Version of Existing Games.” *Procedia – Social and Behavioral Sciences* 191: 487–94. https://doi.org/10.1016/j.sbspro.2015.04.380.

Klemke, Roland, Peter van Rosmalen, Stefaan Ternier, and Wim Westera. 2015. “Keep it Simple: Lowering the Barrier for Authoring Serious Games.” *Simulation and Gaming* 46 (1): 40–67. https://doi.org/10.1177/1046878115591249.

Kultima, Annakaisa. 2015. “Game Design Research.” In *Proceedings of the 19th International Academic Mindtrek Conference*, 18–25. https://doi.org/10.1145/2818187.2818300.

Ladd, Brian, and Ed Harcourt. 2005. “Student Competitions and Bots in an Introductory Programming Course.” *Journal of Computing Sciences in Colleges* 20 (5): 274–84.

Lallemand, Carine, Guillaume Gronier, and Vincent Koenig. 2015. “User Experience: A Concept without Consensus? Exploring Practitioners’ Perspectives through an International Survey.” *Computers in Human Behavior* 43: 35–48. https://doi.org/10.1016/j.chb.2014.10.048.

Lyulina, Elena, Anastasiia Birillo, Vladimir Kovalenko, and Timofey Bryksin. 2021. "TaskTracker-Tool: A Toolkit for Tracking of Code Snapshots and Activity Data during Solution of Programming Tasks." In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, 495–501. https://doi.org/10.1145/3408877.3432534.

Major, Louis, Theocharis Kyriacou, and Pearl Brereton. 2011. "Experiences of Prospective High School Teachers Using a Programming Teaching Tool." In *Proceedings of the 11th Koli Calling International Conference on Computing Education Research*, 126–131. https://doi.org/10.1145/2094131.2094161.

Malik, Sohail Iqbal, and Jo Coldwell-Neilson. 2017. "A Model for Teaching an Introductory Programming Course Using ADRI." *Education and Information Technologies* 22 (3): 1089–120. https://doi.org/10.1007/s10639-016-9474-0.

Malliarakis C., M. Satratzemi. 2014. "Educational Games for Teaching Computer Programming." In *Research on e-Learning and ICT in Education*. Springer, 87–98. https://doi.org/10.1007/978-1-4614-6501-0_7.

Marchiori, Eugenio J., Javier Torrente, Ángel del Blanco, Pablo Moreno-Ger, Pilar Sancho, and Baltasar Fernández-Manjón. 2012. "A Narrative Metaphor to Facilitate Educational Game Authoring." *Computers and Education* 58 (1): 590–99. https://doi.org/10.1016/j.compedu.2011.09.017.

Marti, Patrizia, and Iolanda Iacono. 2015. "Evaluating the Experience of Use of a Squeezable Interface." In *Proceedings of the 11th Biannual Conference on Italian SIGCHI Chapter*. 42–49. https://doi.org/10.1145/2808435.2808461.

Medeiros, R. P., G. L. Ramalho, and T. P. Falcão. 2019. "A Systematic Literature Review on Teaching and Learning Introductory Programming in Higher Education." *IEEE Transactions on Education* 62 (2): 77–90. https://doi.org/10.1109/TE.2018.2864133.

Miller, Craig S., Amber Settle, and John Lalor. 2015. "Learning Object-Oriented Programming in Python: Towards an Inventory of Difficulties and Testing Pitfalls." In *Proceedings of the 16th Annual Conference on Information Technology Education*, Chicago, Illinois, 59–64. https://doi.org/10.1145/2808006.2808017.

Mirel, Barbara. 2004. *Interaction Design for Complex Problem Solving: Developing Useful and Usable Software*. Morgan Kaufmann. https://doi.org/10.1016/B978-155860831-3/50000-X.

Murray, Tom. 2004. "Design Tradeoffs in Usability and Power for Advanced Educational Software Authoring Tools." *Educational Technology* 44 (5): 10–16.

Murray, Tom. 2016. "Coordinating the Complexity of Tools, Tasks, and Users: On Theory-Based Approaches to Authoring Tool Usability." *International Journal of Artificial Intelligence in Education* 26 (1): 37–71. https://doi.org/10.1007/s40593-015-0076-6.

Ni, Lijun. 2009. "What Makes CS Teachers Change?: Factors Influencing CS Teachers' Adoption of Curriculum Innovations." In *SIGCSE '09: Proceedings of the 40th ACM technical symposium on computer science education*, New York, 544–48. https://doi.org/10.1145/1508865.1509051.

Novak, Jasminko, and Susanne Schmidt. 2009. "When Joy Matters: The Importance of Hedonic Stimulation in Collocated Collaboration with Large-Displays." In *IFIP Conference on Human–Computer Interaction*, 618–29. https://doi.org/10.1007/978-3-642-03658-3_66.

Osborn, Joseph C., Melanie Dickinson, Barrett Anderson, Adam Summerville, Jill Denner, David Torres, Noah Wardrip-Fruin, and Michael Mateas. 2019. “Is Your Game Generator Working? Evaluating Gemini, an Intentional Generator.” In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 15: 59–65.

Papastergiou, Marina. 2009. “Digital Game-Based Learning in High School Computer Science Education: Impact on Educational Effectiveness and Student Motivation.” *Computers and Education* 52 (1): 1–12. https://doi.org/10.1016/j.compedu.2008.06.004.

Pérez-Colado, V. M., I. J. Pérez-Colado, M. Freire-Morán, I. Martínez-Ortiz, and B. Fernández-Manjón. 2019. “uAdventure: Simplifying Narrative Serious Games Development.” In *2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT)*, 2161–377X: 119–23. https://doi.org/10.1109/ICALT.2019.00030.

Saavedra, Arturo Barajas, Francisco J. Álvarez Rodríguez, Jaime Muñoz Arteaga, René Santaolaya Salgado, and César A. Collazos Ordoñez. 2014. “A Serious Game Development Process Using Competency Approach: Case Study: Elementary School Math.” In *Proceedings of the XV International Conference on Human Computer Interaction*, Puerto de la Cruz, Tenerife, Spain. https://doi.org/10.1145/2662253.2662352.

Scholtz, Tamarisk Lurlyn, and Ian Sanders. 2010. “Mental Models of Recursion: Investigating Students’ Understanding of Recursion.” In *Proceedings of the fifteenth annual conference on innovation and technology in computer science education*, 103–7. https://doi.org/10.1145/1822090.1822120.

Soloway, Elliot, Mark Guzdial, and Kenneth E. Hay. 1994. “Learner-Centered Design: The Challenge for HCI in the 21st Century.” *Interactions* 1 (2): 36–48. https://doi.org/10.1145/174809.174813.

Tang, Stephen, and Martin Hanneghan. 2010. “A Model-Driven Framework to Support Development of Serious Games for Game-Based Learning. In *Developments in E-systems Engineering (DESE)*, 95–100. https://doi.org/10.1109/DeSE.2010.23.

Tornero, Roberto, Javier Torrente, Pablo Moreno-Ger, and Baltasar Fernández Manjón. 2010. “E-Training DS: An Authoring Tool for Integrating Portable Computer Science Games in E-Learning.” In *International Conference on Web-Based Learning*, 259–68. https://doi.org/10.1007/978-3-642-17407-0_27.

Torrente, Javier, Ángel del Blanco, Guillermo Cañizal, Pablo Moreno-Ger, and Baltasar Fernández-Manjón. 2008. “E-Adventure3D: An Open Source Authoring Environment for 3D Adventure Games in Education.” In *Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology*, Yokohama, Japan, 191–94. https://doi.org/10.1145/1501750.1501795.

Torrente, Javier, Ángel Del Blanco, Eugenio J. Marchiori, Pablo Moreno-Ger, and Baltasar Fernández-Manjón. 2010. “<e-Adventure>: Introducing Educational Games in the Learning Process.” In *IEEE EDUCON 2010 Conference*, 1121–1126. https://doi.org/10.1109/EDUCON.2010.5493056.

Weir, George R. S., Tamar Vilner, António José Mendes, and Marie Nordström. 2005. “Difficulties Teaching Java in CS1 and How We Aim to Solve Them.” *SIGCSE Bulletin* 37 (3): 344–45. https://doi.org/10.1145/1151954.1067543.

Wicentowski, Richard, and Tia Newhall. 2005. “Using Image Processing Projects to Teach CS1 Topics.” *SIGCSE Bulletin* 37 (1): 287–91. https://doi.org/10.1145/1047124.1047445.

Willson, Rebekah. 2019. “Analysing Qualitative Data: You Asked Them, Now What to Do with What They Said.” In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, Glasgow, Scotland 385–87. https://doi.org/10.1145/3295750.3298964.

Barriers to Collaboration in Big Data Analytics Work in Organisations

Mpumelelo Dhlamini
https://orcid.org/0000-0003-2219-9818
Department of Information Systems,
University of Cape Town, South Africa
dhlmpu002@myuct.ac.za

Irwin Brown
https://orcid.org/0000-0001-5321-0771
Department of Information Systems,
University of Cape Town, South Africa
irwin.brown@uct.ac.za

Grant Oosterwyk
https://orcid.org/0000-0002-2745-3929
Department of Information Systems,
University of Cape Town, South Africa
grant.oosterwyk@uct.ac.za

Abstract

With the rise of big data, concepts such as big data analytics were conceived as enablers for data processing to gain valuable insights. The implementation of big data analytics is enterprise-wide and it therefore presents an opportunity for collaboration in analytics work in organisations. Much of the work done with big data analytics, however, is still at departmental level with interdepartmental collaboration often lacking. The purpose of this study is to explore and identify the barriers to collaboration in big data analytics work in organisations. To achieve the research purpose, a qualitative semi-structured interview research strategy was adopted. The data collection was guided by the findings from the literature on the barriers to collaboration in big data analytics work in organisations. Interviews were conducted with 12 IT and analytics professionals from several organisations. A thematic analytic technique was adopted for this study and NVivo was employed to facilitate the analysis process. The barriers to collaboration in big data analytics work identified include activity barriers, capability barriers, context barriers, process barriers, individual professional barriers, team barriers, and technological barriers. A proposed model was developed that depicts these barriers to collaboration in big data analytics work in organisations. The proposed model may serve as a basis for future research.

CCS concepts: information systems, information systems applications, decision support systems, data analytics
Keywords: big data analytics, big data analytics work, collaboration, barriers, organisations

Introduction

Organisations are increasingly producing and harvesting various types of data that are high in volume, velocity and variety, i.e. big data (Behmann and Wu 2015; Walker and Brown 2019). Practitioners across numerous business domains and fields, such as product development, marketing and sales, procurement, customer service, information systems and technology, make use of analytics to obtain insights from this big data (Dremel 2017; Muller et al. 2019; Russom 2011; Walker and Brown 2019). The analytics work makes use of techniques such as descriptive analytics, prescriptive analytics, and predictive analytics (Lepenioti et al. 2020). Big data analytics (BDA) is holistically characterised as "the infrastructure, technologies, tools, methods, techniques and processes used to source, store and analyse big data to produce actionable insights" (Walker and Brown 2019). An organisation-wide BDA competence presents an opportunity for cross-departmental and cross-functional collaboration in BDA work. Collaboration in BDA work may yield great benefits (Akhtar et al. 2019; De Koker 2019; Vera-Baquero, Colomo-Palacios, and Molloy 2013).

Wood and Gray (1991) view collaboration as "a process through which parties who see different aspects of the problem can constructively explore their differences and search for solutions that go beyond their limited vision of what is possible". De Koker (2019) asserts that with the collaboration of various stakeholders, value creation is gained. Even though BDA implementation is enterprise-wide and provides opportunities for collaboration (Heizenberg et al. 2020; Lenz, Wuest, and Westkämper 2018), much of the work done with BDA is at the departmental level, with limited interdepartmental collaboration (Lenz, Wuest, and Westkämper 2018; Russom 2011). As a result, opportunities for knowledge sharing are lost and valuable and reusable resources that could contribute to value creation are wasted (Lenz, Wuest, and Westkämper 2018). Furthermore, Heizenberg et al. (2020) assert that, even with organisations empowering cross-functional teams, maximising data and analytics value, and valuing data and analytics collaboration, collaboration does not appear to occur. There seems to be limits to cross-disciplinary and interdepartmental collaboration for some organisations that are adopting BDA (Dremel 2017; Malaka and Brown 2015).

Lenz, Wuest and Westkämper (2018) suggest that research is needed to develop an assessment tool to help to identify potential barriers that may prove problematic when implementing collaboration in BDA work in an organisation. Such a tool will enable collaboration and better use of the organisation's resources for value creation. To this end, although much research has been done in the various functions of BDA, little is known about collaboration in BDA work in organisations. The purpose of this study is to explore and identify the barriers to collaboration in BDA work in organisations.

Literature Review

Big Data Analytics Work in Organisations

A data-driven organisation is one with multifaceted undertakings that are carried out by different practitioners (Mikalef et al. 2018). BDA work takes place in different organisational activities and domains such as marketing and sales work, product development work and business process analytics (France and Ghose 2019; Grover et al. 2018; LaValle et al. 2011). Collectively, these domains contribute to the business value. Table 1 highlights this interrelated BDA work across the aforementioned business units.

Table 1: Big data analytics work

Big Data Analytics Work	Type of Analytics Work
Marketing and Sales Work	Providing after-sales customer service (Chakravaram, Srinivas, and Ratnakaram 2019). Improving customer relationships, predicting customer behaviour, and predicting product or service sales (France and Ghose 2019). Retaining customers, identifying similar customers (Grover et al. 2018).
Product Development Work	Conducting social media analytics (Rathore and Ilavarasan 2020). Improving business relationships and providing business and social opportunities (Wang et al. 2020). Providing actuarial work (Qin 2020). Identifying new customers (Stephen, Kowolenko, and Michaelis 2015). Assessing product and service performance (Grover et al. 2018).
Business Process Analytics Work	Providing process mining, process discovery, compliance checks, process modelling, and process improvement (Tax et al. 2016). Conducting data analytics and modelling (Sakr et al. 2018). Raising situational awareness and conducting data breach detection, fraud detection, process validation, process verification, process performance, inductive mining, event detection and analysis, bottleneck analysis, and deviation analysis (Grover et al. 2018).

Collaboration in BDA work may benefit an organisation. For example, Kock (2005) notes that business processes can be improved by means of collaboration. The next

section therefore highlights the barriers to collaboration in BDA work in organisations that were identified from the literature.

Barriers to Collaboration in Big Data Analytics Work in Organisations

How collaboration and the factors that influence it are conceptualised is diverse (D'Amour et al. 2005). Since collaboration in a team or alliance is context-specific and affected by its nature, the organisational structure, the participants, the scale of potential conflicts, and the team's or alliance's magnitude (Oraee et al. 2019), it is important to understand the activity, context, motivation and the technology that influence collaboration (Drakos and Gotta 2016).

Drakos and Gotta (2016) note that when considering firstly an activity, for every collaboration initiative there should exist a specific requirement for collaboration, making it necessary to design the collaborative initiative around the requirement. Woodland and Hutton (2012) highlight that a shared purpose is a fundamental requirement to collaboration. Consequently, if the purpose is not recognised, it may impede collaboration. Secondly, there is often a common context between teams or groups that should be highlighted and if it does not exist it should be created. The common context is, for example, a transaction context, an interpersonal relationship context, a group context or a culture context. Thirdly, collaboration requires some form of management to be able to boost the motivation to collaborate; this is through making participation relevant to each individual and the social mechanism appealing to each individual. Lastly, a fit-for-purpose technology for collaboration has potential to enhance collaboration.

Oraee et al. (2019) conceptualised barriers to collaboration in the following categories: (1) process – process management, support tools and technology challenges; (2) context – the environment, organisational and cultural challenges; (3) actor – the knowledge, skills and skills challenges of the participant; (4) team – the team composition, relationship, and knowledge sharing challenges; and (5) task – the structure and demand challenges. Poirier, Forgues and Staub-French (2016) similarly note that collaboration is influenced by the relational structure or system, process, task or activity, and context. Our synthesis of literature resulted in identifying six major barriers, namely, (1) work activity barriers, (2) context barriers, (3) team barriers, (4) process barriers, (5) technological barriers, and (6) individual barriers.

Work Activity Barriers

Not every BDA work activity allows for easy collaboration (Fernandez, Subramaniam, and Franklin 2020). Professionals might feel it is time-consuming, they may be fearful of leaking confidential information, or they may simply not see the benefits (Fernandez, Subramaniam, and Franklin 2020; Kache and Seuring 2017). A lack of consensus on objectives may make professionals' collaboration in a work activity difficult (De Koker 2019; Kache and Seuring 2017).

Context Barriers

The context in which collaboration occurs has an influence on the way in which collaboration occurs in BDA work. Organisational culture and structure are at the heart of whether collaboration in BDA work is enabled or hindered (Bolman and Deal 2017; De Koker 2019; Kaya 2019; Morgan 2015). Organisational structures seem to be influenced by the "organisation's circumstances, including its goals, strategy, technology, and environment" (Bolman and Deal 2017).

Team Barriers

Collaboration in BDA work may be impeded by the way in which collaborating teams are organised and interact with one another. Trust influences the relationship that teams have with one another and has the potential to impede collaboration in BDA work (De Koker 2019). Trust is often breached when there is a violation of an alliance agreement or misconduct (Van den Broek and Van Veenstra 2015). A geographically dispersed team exacerbates the lack of trust, resulting in hindered collaboration (Morrison-Smith and Ruiz 2020). Trust may therefore need to be cultivated for teams to feel comfortable collaborating on data sharing (De Koker 2019). Enforcing interventions such as conflict resolution contracts, cooperation control measures, accountability measures, and data sharing procedures can also be beneficial (Walker and Brown 2019). Many other issues related to geographically dispersed teams, such as time-related, cultural, and linguistic differences, might contribute to collaboration barriers (Morrison-Smith and Ruiz 2020). Communication, cohesion, work norms, mutual support, coordination, and conflict resolution procedures are characteristics that can enable effective collaboration (Hernández 2019). Furthermore, with a well-established collaboration mechanism in place in BDA work, teams can enjoy long-term competitiveness by fostering trust, increasing satisfaction, and sharing information (Akhtar et al. 2019). As a result, they will strengthen their relationships with one another and with their customers. They will be able to better manage their relationships. A joint inventory of partner collaboration, for example, provides a wealth of information that may be used for analysis and forecasting (Akhtar et al. 2019).

Process Barriers

Processes are at the heart of BDA work activities and are often supported and governed by a variety of factors (Grover et al. 2018). Organisations may need to consider the various factors that support and govern collaboration processes for effective collaboration in BDA work. Concerns about privacy, ethics, access, and governance are common in big data and BDA alliances and their processes (Chen, Chiang, and Storey 2012; Daniel 2019). As a result, effective collaboration is hindered. Kache and Seuring (2017) affirm this, stating that although cross-functional integration and collaboration approaches are key, governance and compliance are sometimes barriers to integration and collaboration in BDA work. Furthermore, Mehta, Pandit and Kulkarni (2020) argue that management issues often contribute to process barriers to collaboration in BDA work.

Technological Barriers

Organisations have progressively embraced technologies and tools to enhance collaboration (De Koker 2019). Developers, data owners and data scientists can work together by leveraging collaborative BDA platforms, sharing data, algorithms and services (Park, Nguyen, and Won 2015). With the advancement of mobile devices, social networks and the internet of things, organisations might also seek to transform their analytical processes (Taylor-Sakyi 2016). As a result, numerous companies have adopted collaborative BDA platforms. This is also intended for the purpose of data security (Behmann and Wu 2015). Collaborative BDA platforms maintain data protection and allow various stakeholders to collaborate in their analytics work. Nevertheless, this is still a field that needs further research (Akhtar et al. 2019; Vera-Baquero, Colomo-Palacios, and Molloy 2013). Even with the advantages that technology platforms bring, technologies have drawbacks that can prove to be barriers to collaboration. Technologies that are not fit for purpose in BDA work contribute to the barriers to collaboration (Drakos and Gotta 2016; Morrison-Smith and Ruiz 2020).

Individual Barriers

People who are involved in a collaborative activity are likely to contribute to the barriers to collaboration in BDA work in organisations. This could be owing to issues such as a lack of knowledge of the field domains (Kulkarni et al. 2020; Mikalef et al. 2018). The liaison person's field domain knowledge is critical to facilitating effective collaboration in BDA work (Mikalef et al. 2018). Furthermore, BDA skills are critical to effective collaboration (Ghasemaghaei 2020). Opportunities to leverage collaboration and the organisation's data may be missed if the necessary knowledge and skills are not developed (Kulkarni et al. 2020; Phillips 2017).

Research Design and Methodology

This exploratory research study adopted a qualitative semi-structured interview research strategy. The choice of the research strategy was driven by the desire to increase sample heterogeneity and enhance generalisability. Since the research sought to understand and interpret a social construct (Saunders, Lewis, and Thornhill 2016), the study followed an interpretivist research philosophy and an inductive research approach to obtain new insights, patterns, and a deeper understanding of the fundamentals of the proposed subject. The data collection was guided by the findings from the literature on the barriers to collaboration in BDA work in organisations. Interviews were conducted with IT and analytics professionals involved in BDA work in organisations. The professionals were interviewed in their personal capacity, regardless of their organisational affiliation.

A snowball sampling strategy was used to determine the sample. This was to ensure that knowledgeable professionals are interviewed. Also, the potential of biasness is eliminated (Ishak, Bakar, and Yazid 2014). The sample consisted of 12 IT and analytics professionals. Table 2 presents the demographics of the participants. Most of the

participants were from the financial services industry, and their years of experience ranged from 3 to 20 years, with the majority in senior positions.

Table 2: Demographics and list of the participants in this study

Alias	Participant's Role	Years of Work Experience	Industry
P01S12	Enterprise Architect Head	20	Financial Investment Services
P02S13	IT Operations Head	11	Financial Investment Services
P03S16	Solutions Delivery Head	21	Financial Investment Services
P04S12	Head of Business Intelligence	14	Financial Investment Services
P05S12	Chief Information Officer	16	Financial Investment Services
P06O11	Enterprise Data Head	21	Financial Investment Services
P07O12	Head of Business Intelligence	22	Financial Insurance Services
P08O11	Analyst	7	Financial Investment Services
P09O18	Actuarial Analyst	3	Consultancy
P10O12	Chief Technology Officer	6	Consultancy
P11O16	Data Solutions Manager	14	Retail
P12O10	Head of Data	19	Financial Services

The interview sessions were kept to a maximum of an hour. Interview questions were developed and formed part of the research instrument for data collection. The interview sessions were requested and scheduled at the agreed time, convenient to the participants. The interviews were conducted through the online virtual platform Microsoft Teams. To ensure that ethics in research requirements were met, a consent letter was sent to the participants and consent was obtained before conducting the interviews. Also, a consent letter was sent to the organisation where it was necessary to obtain such consent first, and consent was obtained. Before the interviews, all the participants were informed of what the research was about and that participation was voluntary.

Data Analysis and Findings

Data were collected about the characteristics of BDA work in organisations, the type of collaborative work, and the level of collaboration in the BDA work in organisations. Table 3 presents the findings that emerged from the participants concerning these three dimensions.

Table 3: Big data analytics work

Dimension	Findings	Reference Extracts
Characteristics of BDA work in organisations	• Analytical model development • Data sharing • Data collection • Data curation • Data cleansing • Reporting • Requirements analysis • Data analysis	". . . get insight and find key initiatives . . . make discoveries around data." (P01S12) ". . . you know we want to be a data driven organisation and ensure that . . . data is accessible for quick decision-making and strategic decision making." (P05S12)
Type of collaboration in BDA work in organisations	• Cross-functional collaboration • Interdepartmental collaboration • Cross-business-unit collaboration	". . . typically not cross department . . . the data engineering team is like one team." (P10O12) ". . . I have a team that reports to me who run the big data environment. We then have the analytics and optimisation team. And there's a couple of . . . who need to make utilisation of the big data environment and the data that's stored there for analytical operations." (P12O10)
Level of collaboration in BDA work in organisations	• Formal: contractual • Informal: ad hoc, social contract	". . . whether is formal or informal, we have to be very clear . . ." (P06O11)

Barriers to Collaboration in BDA Work in Organisations

Seven major barriers to collaboration in BDA work in organisations were identified, namely, (1) activity barriers, (2) capability barriers, (3) context barriers, (4) process barriers, (5) individual professional barriers, (6) team barriers, and (7) technological barriers.

Activity Barriers

Almost all participants shared the sentiment that the lack of shared understanding, interest, objectives, and goals limit collaboration in BDA work activities. It was noted that

> . . . if that shared vision isn't there . . . it's really difficult to achieve those set objectives. (P05S12)

> . . . the business struggles to sort of understand you know what they want. (P03S16)

Time-related barriers, which included priority issues and workloads, were noted as impeding collaboration in BDA work activities. For example, three participants argued that

> Some business units are/have different timing needs. (P11O16)

> It will sort of waste time because you're trying to fix data that should have been cleaned. (P09O18)

> There's a lot of work coming through from multiple parties. (P12O10)

A few participants highlighted barriers associated with trust in the BDA work activities:

> . . . you don't want to, you know, just give anybody . . . the minute you change a rule . . . it can be the downfall of an organisation. (P11O16)

> . . . obviously the trust from the team members themselves. Because now we know that so and so is underperforming. (P10O12)

> . . . it will affect collaboration . . . we can't just continue and trust the data set and continue. (P09O18)

Collaboration is hampered when professionals engage in an activity that is less of a priority and its contribution to business value is not immediate. A participant stated: "In the informal collaboration . . . the biggest challenge is getting innovative ideas to become a reality . . . nothing really gets prioritised" (P05S12).

Capability Barriers

Most of the participants highlighted the lack of maturity in the capability of BDA as a barrier to collaboration in BDA work in organisations. One participant noted, "it would be that we haven't really matured the data platforms that BI and analytics . . ." (P02S13). The participants also pointed out that collaboration in the BDA work is constrained since the full capabilities of the big data stack are not being leveraged. One participant stated: "you're not leveraging the full capabilities of the big data stack" (P12O10).

A few participants argued that even with an established BDA capability, the lack of BDA knowledge contributes to inadequate collaboration in BDA work in organisations. One participant stated: "the lack of knowledge or competency . . . big data being sort of a new thing coming into the industry" (P09O18).

Context Barriers

Most participants noted that the organisational structure might pose barriers to collaboration. One participant stated: "If everybody reported to the same boss . . . from a structural perspective, it might be easier to then collaborate" (P12O10). In addition, the silo operation is seen by other participants as a barrier, as stated by one participant: "the organisational structure is set in such a way . . . operate in silos" (P05S12). Some participants, however, claim that having a decentralised BDA capability has its own advantages and disadvantages. One participant stated:

> . . . maybe there is pros and cons to both . . . for example, . . . decentralised data management capability . . . the minute you decentralise something, there's a lot of parties that now have a voice in terms of . . . governance and decision making around data . . . And now it becomes more and more decentralised, which maybe, to some extent is good, because . . . it's a flat structure. (P11O16)

Most participants argued that organisational culture too contributes to collaboration challenges. The human aspect and the personal qualities of a professional often add to the barriers. One participant stated: "you have instances, where there may be some resistance, but that's the nature of humans I guess" (P01S12). It was also noted that the culture of specialists impedes collaboration in BDA work. One participant stated:

> There are pockets in the business or pockets in the teams, that some of the people would rather be . . . more specialised . . . they pretty much very specialised and locked into what they do. (P02S13)

Process Barriers

Almost all the participants highlighted challenges related to data management as barriers to collaboration. One participant stated: "trying to make the data more accessible and usable, but again, it's also not to everyone . . . allowed to see everything" (P08O11). Most of the participants claimed that regulatory requirements lead to barriers to collaboration in BDA work. One participant noted: "one of the barriers of collaboration and of getting things done quicker, is that with the regulation that come in, especially the POPIA regulation, there is quite a lot of other regulations that control data" (P03S16).

One participant pointed out that inconsistencies in the reporting and representation standards contribute to barriers to collaboration: "you find that you manipulate data in a different way. To a point that whoever takes on your project doesn't even understand the sort of manipulation that was done" (P09O18). The manner in which the

manipulated data is presented may not inherently be desirable owing to conflicting standards. One participant stated: "it becomes a challenge because the format by which my team present or playback the data to the businesses is either undesirable" (P02S13).

It was also noted that the engagement model may lead to barriers to collaboration. One participant stated:

> . . . the construction of your models when it comes to data that you are consuming and so forth, and then running the respective analysis and so forth . . . the engagement model is quite a big a challenge, in my view. (P05S12)

Individual Professional Barriers

A few participants argued that concerns regarding the level of professional competence contribute to barriers to collaboration in BDA work in organisations. For example, one participant noted:

> . . . them converting the model to an API, the data scientist is not really familiar with . . . and then they rely on the engineers to be able to assist him with that task . . . people are not willing to step out of their comfort zone to learn new things. (P10O12)

It emerged that the lack of both domain knowledge and analytics competency has an influence on effective collaboration. One participant stated: "some of the analytics guys don't necessarily understand that environment . . . the analytics guys sometimes don't understand how things work in a big data space" (P12O10).

A few participants highlighted capacity challenges, such as the lack of time and a single point of contact as contributing to barriers to collaboration. One participant stated:

> So even the ownership of or stewardship of data related queries, become something that you would speak to an SME . . . that individual will have their own priorities . . . if the individual is away on leave or off sick we jeopardise . . . (P02S13)

Team Barriers

Some participants noted that team dynamics affects collaboration. One participant stated: "you need to build up a track record that as an analytics competency . . . you're only there to actually try and shift the business forward" (P04S12). A few pointed out that misunderstandings and different jargon have an impact on the effectiveness of collaboration. One participant noted: "when it happens, it is by a request from them for stuff, or we send them back the information, then there seems to be a gap between the understanding, or we don't necessarily understand exactly what they mean" (P12O10).

It was noted the challenges associated with continuous learning and knowledge sharing of the ever-changing BDA landscape have an impact on effective collaboration:

> I think the landscape is moving and shifting such that there's always continuous learning . . . because it's new territory for either teams it becomes a bit uncomfortable . . . when people are coming from different school of thoughts there will obviously be some contention or uncomfortably in that regard. (P02S13)

It was noted that the difficulties that emerge as a result of a team being made up of subject matter experts, with unclear roles and responsibilities, hinder collaboration. One participant stated: "so you'd find that the composition is different in that . . . Some people are more experienced than the other people. And that does affect collaboration" (P09O18).

Technological Barriers

Almost all the participants agreed that analytics technologies contribute to barriers to collaboration. Challenges regarding technology, fit-for-purpose tools, tool complexity, maturity of technology and an unstandardised technology stack are observed as barriers to collaboration in BDA work. One participant stated:

> A system that is just built now. Ah, it's very difficult to trust its output . . . And it might be very difficult to pick it . . . the flip side of that is also mature systems can be very problematic . . . it can't take in the sort of data and the structure of the data that you need in big data. (P09O18)

A few participants highlighted collaborative platforms as contributing to collaboration barriers in BDA work in organisations. It was noted that the unstandardised use of collaborative tools or platforms contributes to barriers to collaboration:

> The current context with majority of us now in the business 90% to 95% are working from home . . . are no longer co-located. You need to depend on written communication . . . sort of move away from synchronous to asynchronous. And as such that becomes a challenge . . . because you now talking to people that we don't have conformance in terms of tooling. (P02S13)

Findings and Implications

Activity barriers such as the lack of shared objectives and understanding, time constraints, lack of trust, and lack of understanding of the significance, value and benefits of collaboration were identified as affecting collaboration in BDA work in organisations. The literature confirms these as barriers (Akhtar et al. 2019; Fernandez, Subramaniam, and Franklin 2020; Kache and Seuring 2017). Barriers related to capability maturity and BDA knowledge emerged too as affecting collaboration in BDA work. Mikalef et al. (2018) argue that in order to exploit big data and big data-related resources to gain insight into data, organisations need to increase their proficiency in BDA capabilities.

Contextual barriers such as organisational structure affect collaboration in BDA work, as identified in the literature (De Koker 2019). Similarly, organisational culture was seen as affecting collaboration, which affirmed the literature (Kaya 2019). Process barriers associated with data governance, regulation and inconsistent data presentation and data quality standards emerged as affecting collaboration in BDA work. The literature too identifies such issues (Behmann and Wu 2015; Chen, Chiang, and Storey 2012).

Team barriers linked to capacity, trust, communication, team composition, continuous learning and knowledge sharing culture also emerged. The literature confirms that a team's culture should promote collaboration by encouraging trust and communication (Akhtar et al. 2019; De Koker 2019; Morrison-Smith and Ruiz 2020).

The technological barriers that emerged included issues related to analytics technologies and tools, and collaboration platform technologies. Standardisation, maturity and the complexity of technologies and tools are the key contributors to collaboration barriers in BDA work. Although technology and tools also enable collaboration, the drawbacks associated with them negatively affect collaboration (Gotta, Preset, and Elliot 2018; Kulkarni et al. 2020; Mikalef et al. 2018).

In summary, although most findings affirm those barriers identified through the thematic literature review, they do so via a synthesised and rich description in the specific context of BDA work in organisations. The findings provide a basis for formulating a proposed model as presented next.

Proposed Model

As described in the preceding sections, the study identified various barriers to collaboration in BDA work in organisations, which served as the foundation for developing the proposed model. Figure 1 depicts the proposed model, which categorises the barriers as activity and context barriers, team barriers, process and technological barriers, individual professional barriers, and capability barriers.

Figure 1: Proposed model for barriers to collaboration in big data analytics work in organisations

Conclusion

The barriers to collaboration in BDA work that emerged from this study include activity barriers, capability barriers, context barriers, process barriers, individual professional barriers, team barriers, and technological barriers. Almost all these barriers were firstly identified through a comprehensive synthesis of the literature. Capability barriers emerged as a new theme through the primary data collection and analysis. A proposed model has been developed that can help organisations understand and possibly develop an evaluation method to identify potential barriers to collaboration in BDA work in organisations.

The study does not explain fully the way in which the barriers identified impede collaboration in BDA work in organisations. Future research should seek to understand

and explain the way in which the barriers to collaboration in BDA work in organisations impede collaboration. Future studies may also seek to understand the way in which barriers to collaboration in BDA work in organisations relate and influence each other. In addition, future studies may seek to use the proposed model to develop an evaluation method to identify potential barriers to collaboration in BDA work in organisations. Furthermore, each identified barrier is broad in scope and is most likely influenced by a variety of underlying factors. Future studies should consider focusing on a specific barrier in greater depth. For example, a multi-case study could examine the way in which different organisational contexts affect collaboration in BDA work.

References

Akhtar, P., Z. Khan, R. Rao-Nicholson, and M. Zhang. 2019. "Building Relationship Innovation in Global Collaborative Partnerships: Big Data Analytics and Traditional Organizational Powers." *R&D Management* 49 (1): 7–20. https://doi.org/10.1111/radm.12253.

Behmann, F., and K. Wu. 2015. *Collaborative Internet of Things (C-IOT): For Future Smart Connected Life and Business*. London: John Wiley and Sons. https://doi.org/10.1002/9781118913734.

Bolman, L. G., and T. E. Deal. 2017. *Reframing Organizations: Artistry, Choice, and Leadership*. John Wiley and Sons. https://doi.org/10.1002/9781119281856.

Chakravaram, V., J. Srinivas, and S. Ratnakaram. 2019. "The Role of Big Data, Data Science and Data Analytics in Financial Engineering." In *Proceedings of the 2019 International Conference on Big Data Engineering*, 44–50. https://doi.org/10.1145/3341620.3341630.

Chen, H., R. H. Chiang, and V. C. Storey. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact." *MIS Quarterly* 36 (4): 1165–88. https://doi.org/10.2307/41703503.

D'Amour, D., M. Ferrada-Videla, L. San Martin Rodriguez, and M. D. Beaulieu. 2005. "The Conceptual Basis for Interprofessional Collaboration: Core Concepts and Theoretical Frameworks." *Journal of Interprofessional Care* 19 (sup1): 116–31. https://doi.org/10.1080/13561820500082529.

Daniel, B. K. 2019. "Big Data and Data Science: A Critical Review of Issues for Educational Research." *British Journal of Educational Technology* 50 (1): 101–13. https://doi.org/10.1111/bjet.12595.

De Koker, L. 2019. "Fostering Collaboration amongst Business Intelligence, Business Decision Makers and Statisticians for the Optimal Use of Big Data in Marketing Strategies." PhD thesis, University of the Western Cape.

Drakos, N., and M. Gotta. 2016. "How to Make Collaboration Work with Gartner's ACME Framework." Gartner. Accessed 14 October 2021. https://www.gartner.com/en/documents/3331317/how-to-make-collaboration-work-with-gartner-s-acme-frame.

Dremel, C. 2017. "Barriers to the Adoption of Big Data Analytics in the Automotive Sector." In *Proceedings of AMCIS*, Boston, MA, USA.

Fernandez, R. C., P. Subramaniam, and M. J. Franklin. 2020. "Data Market Platforms: Trading Data Assets to Solve Data Problems." *Proceedings of the VLDB Endowment* 13 (11): 1–15. https://doi.org/10.14778/3407790.3407800.

France, S. L., and S. Ghose. 2019. “Marketing Analytics: Methods, Practice, Implementation, and Links to Other Fields.” *Expert Systems with Applications* 119: 456–75. https://doi.org/10.1016/j.eswa.2018.11.002.

Ghasemaghaei, M. 2020. “The Role of Positive and Negative Valence Factors on the Impact of Bigness of Data on Big Data Analytics Usage.” *International Journal of Information Management* 50: 395–404. https://doi.org/10.1016/j.ijinfomgt.2018.12.011.

Gotta, M., A. Preset, and B. Elliot. 2018. “Embrace Workstream Collaboration to Transform Team Coordination and Performance.” Gartner. Accessed 18 October 2021. https://www.gartner.com/en/documents/3712617/embrace-workstream-collaboration-to-transform-team-coord.

Grover, V., R. H. Chiang, T. P. Liang, and D. Zhang. 2018. “Creating Strategic Business Value from Big Data Analytics: A Research Framework.” *Journal of Management Information Systems* 35 (2): 388–423. https://doi.org/10.1080/07421222.2018.1451951.

Heizenberg, J., K. Schlegel, F. Buytendijk, and A. Kronz. 2020. “Create a Hybrid Centralized and Decentralized Data and Analytics Organizational Model.” Gartner. Accessed 12 October 2021. https://www.gartner.com/en/documents/3980295/create-a-hybrid-centralized-and-decentralized-data-and-a.

Hernández, A. K. L. 2019. “Team Collaboration Capabilities as Drivers for Innovation Performance: The Case of Spanish Technology-Based Startups.” PhD dissertation, University of Valencia.

Ishak, N. M., A. Bakar, and A. Yazid. 2014. “Developing Sampling Frame for Case Study: Challenges and Conditions.” *World Journal of Education* 4 (3): 29–35. https://doi.org/10.5430/wje.v4n3p29.

Kache, F., and S. Seuring. 2017. “Challenges and Opportunities of Digital Information at the Intersection of Big Data Analytics and Supply Chain Management.” *International Journal of Operations and Production Management* 37 (1): 10–36. https://doi.org/10.1108/IJOPM-02-2015-0078.

Kaya, D. 2019. “Intra-Organizational Collaboration for Innovation: Understanding the Dynamics of Formal and Informal Structures.” Master’s thesis, KTH Royal Institute of Technology. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-263170.

Kock, N., ed. 2005. *Business Process Improvement through E-Collaboration: Knowledge Sharing through the Use of Virtual Groups: Knowledge Sharing through the Use of Virtual Groups*. Hershey: IGI Global. https://doi.org/10.4018/978-1-59140-357-9.

Kulkarni, A. J., P. Siarry, P. K. Singh, A. Abraham, M. Zhang, A. Zomaya, and F. Baki, eds. 2020. *Big Data Analytics in Healthcare*. Springer. https://doi.org/10.1007/978-3-030-31672-3.

LaValle, S., E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz. 2011. “Big Data, Analytics and the Path from Insights to Value.” *MIT Sloan Management Review* 52 (2): 21–32.

Lenz, J., T. Wuest, and E. Westkämper. 2018. “Holistic Approach to Machine Tool Data Analytics.” *Journal of Manufacturing Systems* 48: 180–91. https://doi.org/10.1016/j.jmsy.2018.03.003.

Lepenioti, K., A. Bousdekis, D. Apostolou, and G. Mentzas. 2020. “Prescriptive Analytics: Literature Review and Research Challenges.” *International Journal of Information Management* 50: 57–70. https://doi.org/10.1016/j.ijinfomgt.2019.04.003.

Malaka, I., and I. Brown. 2015. “Challenges to the Organisational Adoption of Big Data Analytics: A Case Study in the South African Telecommunications Industry.” In *Proceedings of the 2015 annual research conference on South African Institute of Computer Scientists and Information Technologists* 27: 1–9. https://doi.org/10.1145/2815782.2815793.

Mehta, N., A. Pandit, and M. Kulkarni. 2020. “Elements of Healthcare Big Data Analytics.” In *Big Data Analytics in Healthcare*, 24–43. Cham: Springer. https://doi.org/10.1007/978-3-030-31672-3_2.

Mikalef, P., I. O. Pappas, J. Krogstie, and M. Giannakos. 2018. “Big Data Analytics Capabilities: A Systematic Literature Review and Research Agenda.” *Information Systems and e-Business Management* 16 (3): 547–78. https://doi.org/10.1007/s10257-017-0362-y.

Morgan, J. 2015. “The 5 Types of Organizational Structures: Part 1, The Hierarchy.” Accessed 14 October 2021. https://www.forbes.com/sites/jacobmorgan/2015/07/06/the-5-types-of-organizational-structures-part-1-the-hierarchy/amp.

Morrison-Smith, S., and J. Ruiz. 2020. “Challenges and Barriers in Virtual Teams: A Literature Review.” *SN Applied Sciences* 2: 1–33. https://doi.org/10.1007/s42452-020-2801-5.

Muller, M., L. Lange, D. Wang, D. Piorkowski, J. Tsay, Q. V. Liao, and T. Erickson. 2019. “How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation.” In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3290605.3300356.

Oraee, M., M. R. Hosseini, D. J. Edwards, H. Li, E. Papadonikolaki, and D. Cao. 2019. “Collaboration Barriers in BIM-Based Construction Networks: A Conceptual Model.” *International Journal of Project Management* 37 (6): 839–54. https://doi.org/10.1016/j.ijproman.2019.05.004.

Park, K., M. C. Nguyen, and H. Won. 2015. “Web-Based Collaborative Big Data Analytics on Big Data as a Service Platform.” In *2015 17th international conference on advanced communication technology,* 564–567. IEEE. https://doi.org/10.1109/ICACT.2015.7224859.

Phillips, F. 2017. “A Perspective on ‘Big Data’.” *Science and Public Policy* 44 (5): 730–37. https://doi.org/10.1093/scipol/scx012.

Poirier, E., D. Forgues, and S. Staub-French. 2016. “Collaboration through Innovation: Implications for Expertise in the AEC Sector.” *Construction Management and Economics* 34 (11): 769–89. https://doi.org/10.1080/01446193.2016.1206660.

Qin, Y. 2020. “The Use of IT in Life and Health Insurance Product Development.” Master’s thesis, Åbo Akademi University.

Rathore, A. K., and P. V. Ilavarasan. 2020. “Pre- and Post-Launch Emotions in New Product Development: Insights from Twitter Analytics of Three Products.” *International Journal of Information Management* 50: 111–27. https://doi.org/10.1016/j.ijinfomgt.2019.05.015.

Russom, P. 2011. *Big Data Analytics. TDWI Best Practices Report*. Renton: TDWI.

Sakr, S., Z. Maamar, A. Awad, B. Benatallah, and W. M. van der Aalst. 2018. “Business Process Analytics and Big Data Systems: A Roadmap to Bridge the Gap.” *IEEE Access* 6: 77308–20. https://doi.org/10.1109/ACCESS.2018.2881759.

Saunders, M., P. Lewis, and A. Thornhill. 2016. *Research Methods for Business Students*. 7th ed. New York: Pearson.

Stephen, K. Markham, Michael Kowolenko, and Timothy L. Michaelis. 2015. “Unstructured Text Analytics to Support New Product Development Decisions.” *Research-Technology Management* 58 (2): 30–39. https://doi.org/10.5437/08956308X5802291.

Tax, N., N. Sidorova, R. Haakma, and W. M. van der Aalst. 2016. “Event Abstraction for Process Mining Using Supervised Learning Techniques.” In *Proceedings of SAI Intelligent Systems Conference*, edited by Y. Bi, S. Kapoor and R. Bhatia, 251–269. Cham: Springer.

Taylor-Sakyi, K. 2016. “Big Data: Understanding Big Data.” Cornell University. Accessed 14 October 2021. https://arxiv.org/abs/1601.04602.

Van den Broek, T. A., and A. F. van Veenstra. 2015. “Modes of Governance in Inter-Organizational Data Collaborations.” *ECIS 2015 Completed Research Papers*. Paper 188. http://aisel.aisnet.org/ecis2015_cr/188.

Vera-Baquero, A., R. Colomo-Palacios, and O. Molloy. 2013. “Business Process Analytics Using a Big Data Approach.” *IT Professional* 15 (6): 29–35. https://doi.org/10.1109/MITP.2013.60.

Walker, R. S., and I. Brown. 2019. “Big Data Analytics Adoption: A Case Study in a Large South African Telecommunications Organisation.” *South African Journal of Information Management* 21 (1): a1079. https://doi.org/10.4102/sajim.v21i1.1079.

Wang, Y., M. Rod, Q. Deng, and S. Ji. 2020. “Exploiting Business Networks in the Age of Social Media: The Use and Integration of Social Media Analytics in B2B Marketing.” *Journal of Business and Industrial Marketing*. https://doi.org/10.1108/JBIM-05-2019-0173.

Wood, D. J., and B. Gray, B. 1991. “Toward a Comprehensive Theory of Collaboration.” *Journal of Applied Behavioral Science* 27 (2): 139–62. https://doi.org/10.1177/0021886391272001.

Woodland, R. H., and M. S. Hutton. 2012. “Evaluating Organizational Collaborations: Suggested Entry Points and Strategies.” *American Journal of Evaluation* 33 (3): 366–83. https://doi.org/10.1177/1098214012440028.