2021

A Comparison of Instructional Efficiency Models in Third Level Education

Murali Rajendran
Technological University Dublin

Recommended Citation

Rajendran, M. (2021). A Comparison of Instructional Efficiency Models in Third Level Education. Technological University Dublin. DOI: 10.21427/YRCX-JW95

A Comparison of Instructional Efficiency Models in Third Level Education

Murali Rajendran

A dissertation submitted in partial fulfilment of the requirements of
Technological University Dublin for the degree of
M.Sc. in Computer Science (Data Analytics)

02 March 2021

Declaration

I certify that this dissertation which I now submit for examination for the award of M.Sc. in Computer Science (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technological University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: *Murali Rajendran*

Date: *02 March 2021*

Abstract

This study investigates the validity and sensitivity of a novel model of instructional efficiency: the parabolic model. The novel model is compared against state-of-the-art models present in instructional design today; Likelihood model, Deviational model and Multidimensional model. This models is based on the assumption that optimal mental workload and high performance leads to high efficiency, while other models assume that low mental workload and high performance leads to high efficiency. The investigation makes use of two instructional design conditions: a direct instructions approach to learning and its extension with a collaborative activity. A control group received the former instructional design while an experimental group received the latter design. A performance score was extracted for evaluation. The models of efficiency compared were based upon both a unidimensional and a multidimensional measure of mental workload, which were acquired through self-reporting from the participants. These mental load measures in conjunction with the performance score contribute to the calculation of efficiency scores for each model. The aim of this study is to determine whether the novel model is able to better differentiate between the control and experimental groups based on the resulting efficiency when compared to the other models. The models were analysed and compared using various statistical tests and techniques. Empirical evidence partially supports the proposed hypothesis that parabolic model demonstrates validity, however lacks sufficient statistical evidence to suggest that the model has better sensitivity and its capacity to differentiate between the two groups.

Keywords: Instructional design, Mental workload, Instructional efficiency, Validity, Sensitivity, Shift function, Classification.

Acknowledgments

I would like to express my gratitude to my supervisor Dr. Luca Longo for his continued support, enthusiasm, patience and encouragement. This work would not have been possible without your experience and valuable guidance. Thank you for always being there when needed and pushing me to bring out my best.

I would like to thank Mr. Giuliano Orru for sharing his expertise and useful advice which allowed me to develop this research and expand into the thesis it is today.

I would like to thank the academic staff in TU Dublin for teaching me over the years, allowing me to build a foundation in data analysis which was crucial for my research.

I want to thank my colleagues for their continued encouragement and my manager Mr. Horgan for allowing me the flexibility I required at work to manage my workload efficiently between work and college. For that, I am very grateful.

I would like to thank my friends for keeping me motivated throughout. Your presence created a salutary effect and allowed me to work with much focus.

Finally, my dear mother, words cannot describe what I want to say. Thank you for being my beacon of hope throughout the years. Your love, care, encouragement, and direction has brought me to this very point in life. Without you, none of this is possible.

Contents

List of Figures

List of Tables

List of Acronyms

CLT	Cognitive load theory
DM	Deviational model
HCA	Human cognitive architecture
IG	Information gain
INS.EFF	Instructional efficiency
LM	Likelihood model
LR.EFF	Learning efficiency
MCQ	Multiple-choice questionnaire
MWL	Mental workload
NASA-TLX	National Aeronautics and Space Administration Task Load Index
PM	Parabolic model
RAWNASA	Raw NASA-TLX score
RSME	Rating scale mental effort
SVM	Support vector machine
SMOTE	Synthetic minority over-sampling technique
TR.EFF	Training efficiency
3DM	Multidimensional model

Chapter 1

Introduction

1.1 Background

Cognitive Load Theory is a widely known theory in educational psychology. It assumes that working memory can process only explicit instructions. Explicit instructions are direct and specific explanations aimed at facilitating problem solving or completion of task. This is often referred to as the traditional teaching method and is structured. Another method is the inquiry activity which is aimed at engaging learners by the use of focused communication focused on cognitive trigger questions. This process of forming knowledge is rather ill structured, as the goal is to reach an agreement and construct knowledge collaboratively. Efficiency in learning and instruction is the capacity to achieve established goals with minimum expenditure of effort or resources (Hoffman & Schraw, 2010).

Efficiency is calculated based on the mental effort or workload exerted during a task and the performance outcome. Ideally, any activity conducted should be as efficient as possible. Research have been conducted on teaching methodologies that aims at combining the traditional teaching method and a community of inquiry approach by extending the former with the latter and comparing its efficiency (experimental) versus the efficiency of traditional method alone (control). It is important to understand how particular approaches to learning influence the performance of learners.

1.2 Research Problem

The comparison of groups typically involves the use of one central tendency value, e.g. mean, which are not robust. It assumes that distributions differ only in the central tendency, and not in other aspects. Tests on mean values are not robust to outliers, skewness, heavy-tails, and for independent groups, differences in skewness, variance and combinations of these factors (Wilcox & Keselman, 2003; Wilcox, 2012). Therefore, it is important to consider and compare entire distributions. Considering there are more informative statistics available, there is no particular reason why the mean should be used to compare two different groups.

It has emerged in the past that in education, the assumption / rationale that underpins efficiency is that low mental effort with high performance scores provides the best efficiency. By contrast, high mental effort with low performance provides the worst efficiency. Although the framework of optimal effort / mental workload is applied widely in other fields, it is not widely used when it comes to instructional efficiency. Another problem with the current models of efficiency is that either they are affected by variability of all the observations in the group or that they are sensitive to minor changes in the sample of observations. The parabolic model looks to addresses these issues. The parabolic model assumes that optimal workload and high performance provides the best efficiency. Currently there are no applications of the parabolic model in the field of education.

The current research proposes introducing the parabolic model and comparing it with other models of efficiency, the models can be compared and the model which best differentiates the groups can be determined. The study must be limited to environments that only use the traditional method approach to learning so that the comparison can be justified. There are a number of issues with small sample sizes such as low reproduciblity and inflated effect size estimations which must be addressed where possible.

1.3 Research Objectives

The aim of this research is to evaluate the effectiveness of the parabolic model and determine whether it is able to distinguish between the control and experimental groups better than the existing models of efficiency in third level education.

The objectives of this research are as follows:

1. Conduct a literature review and understand the state-of-the-art knowledge that is present regarding the area of research.
2. Conduct a secondary empirical research.
    - Form the research hypothesis.
    - Explore, understand and process the data.
    - Explain the various methods in the design.
    - Determine the most suitable techniques.
3. Implement the design of the experiment in R and obtain the results.
4. Evaluate the results and discuss the findings to verify whether the research hypothesis is accepted or rejected.
5. Identify and suggest future areas of research which may improve and assist in determining a better understanding of the models of efficiency in third level education.

1.4 Research Methodologies

This work involves summary, collation and analysis of existing research, and the use of pre-existing data makes this a secondary research. The comparative study conducted will be an empirical investigation on quantitative properties of the dataset, which are direct and measurable. It is aimed at formulating and testing a hypothesis pertaining to the features. The results will be plotted on a suitable graphical method for comparison and a logical conclusion will be made from the available facts. The use of empirical evaluation techniques, such as statistical methods and models employed, establishes a deductive basis for future problems.

1.5 Scope, Limitations and Delimitations

The scope of this study is limited to lectures related to Computer Science in Technological University Dublin, that use the in-person traditional direct instruction approach to learning in third level education. Lectures delivered online and labs were outside the scope of this research.

The number of participants from the different lectures varied each time and was inconsistent, leading to inconsistent splitting of the groups and small sample sizes,which means generalisability may not be possible. Another limitation faced is the number of observations in the dataset (455). Due to the ongoing pandemic situation, the in-person collection of data from lectures was no longer possible and had to cease from February 2020.

Various methods of measuring mental workload such as Workload Profile and physiological measures were rejected in favour of subjective measures such as RSME and Raw NASA-TLX, because they were outside the scope of this research.

1.6 Document Outline

Chapter 2: Literature Review informs the reader on the related background that originated this research. It describes the relevant literature related to key areas of this research such as cognitive load theory, instructional design, models of instructional efficiencies and mental workload measures.This chapter will conclude with identifying the research gaps in the existing literature and the research question.

Chapter 3: Design and Methodology describes the design of the empirical experimental framework and the methodology employed with the aim to solve the research question. It will detail the collection of data, the models and statistics to be employed and evaluation of the models. The strengths and limitations of the design of this experiment will also be discussed in this chapter.

Chapter 4: Results, Evaluation and Discussion outlines the results, its analysis and interprets the data from the implementation of the experiment. The results will be analysed and the models will be evaluated as described in **Section 3.5**. The chapter will also discuss strengths and limitations of the results and evaluation from this experiment and problems encountered.

Chapter 5: Conclusion will conclude the research by providing a summary of the work carried out, highlighting the contribution to the general body of research within instructional efficiency in third level education. Further areas of investigation and research will be addressed in order to potentially improve on the results found for future work.

Chapter 2

Literature Review

The aim of this chapter is to inform the reader with basic notions of cognitive load theory, mental workload, instructional design, instructional efficiency and the different models of efficiency used in the third level education setting. This theoretical content is crucial to provide a clear layout of the proposed experiment as devised in chapter 3.

2.1 Cognitive Load Theory

Cognitive Load Theory (CLT) is a widely known theory in Educational Psychology, which is used to enhance the learning phase by developing or applying instructional teaching techniques based on the limitations of human cognitive architecture. This is done by finding the optimum cognitive load imposed on working memory of learners while performing a task. Cognitive Load Theory provides an effective framework for designing and delivering work to learners of any standard. It is backed by empirical research supporting different amounts and types of instruction according to the level of learners and it enables instructors to provide well crafted guidance in their topics. It states that effective learning can only take place where the cognitive capacity of an individual in a particular domain is not exceeded. Human Cognitive Architecture (HCA) provides a generic framework of the information-processing stages that learners use to encode, store, and modify information for the purposes reasoning and decision making (Atkinson & Shiffrin, 1968; Reed, 2012). It describes the necessary and sufficient conditions for a human to input, process and store the data which in turn becomes information and output the results. Cognitive Load includes units of knowledge and elements of relationship. The Cognitive Load of a task is created when the units of knowledge interact with the relationship elements (Sweller, 2003).

Sensory Memory, Short-term Memory, also known as Working Memory, and Long-term Memory are three essential dimensions of HCA. Atkinson and Shiffrin(1968) propose that the input of data entered via sensory memory is processed in the working memory and then proceeds to be stored in the Long-term memory through their Modal model in Human Processing. Working memory is limited and it processes incoming information from sensory memory, long term memory instead is unlimited, highly structured and it stores relevant information as acquired knowledge (Miller, 1956; Baddeley, 2001). Short term memory, as described by Miller(1956), has the capacity to hold seven plus or minus two chunks of information at any given time. It is not specified whether the chunks of information were novel or familiar, interrelated or discrete; simply that a chunk is a unit of knowledge. Long-term memory is a permanent store of experience, knowledge and process, all of which is held outside the conscious awareness until recalled in the working memory. It does not have an executive function (Baddeley, 2001). The information stored in the long-term memory in knowledge structures of varying complexity is called ”Schemata” (Sweller, 2003). These schemata makes the construction and transfer of knowledge possible. This is the goal of learning. The more schemata an individual holds for a particular topic, the more advanced they become in learning. Schema construction is believed to reduce the load in working memory. Leaving sufficient cognitive resources in the working memory to process new information is one of the core objectives of educational instructional design (Orru & Longo, 2019b). Explicit instructions are required to process information and build schemata of knowledge in working memory. Traditionally, cognitive load theory has focused on instructional methods to decrease extraneous cognitive load so that the available resources can be fully devoted to learning. Many methods have been devised and will continue to be devised (Sweller, 2003).

2.1.1 Types of Cognitive Load

Cognitive Load Theory distinguishes three types of cognitive load. They are Intrinsic Load, Extraneous Load and Germane Load (Sweller, 2010; van Gog & Paas, 2008).

Intrinsic Load

Intrinsic Cognitive Load is concerned with the underlying complexity of information which must be understood unencumbered by instructional issues (Sweller, 2010). The intrinsic load is the degree of element interactivity during problem solving. Element interactivity corresponds to the number of information to be simultaneously processed by working memory in problem solving or in task learning (Orru & Longo, 2019b). Materials with a low element interactivity imposes low working memory load because individual elements can be learned with minimal reference to other elements, whereas high element interactivity imposes high working memory load because elements interact with each other heavily and so cannot be learned in isolation.

Every knowledge unit has an intrinsic cognitive load. While it can be reduced to rote components initially, it still has a minimum, irreducible cognitive load (Van Merrienboer & Sweller, 2005). For a long time, intrinsic load was considered unalterable by instruction, but recently some research effort has been devoted to finding techniques to manage this load (Pollock, Chandler, & Sweller, 2002), which may be unavoidable in situations where tasks are extremely complex for learning to commence. Sweller, Van Merrienboer, and Paas (2019) also acknowledge that for a given task and learner's knowledge level, intrinsic load is fixed and cannot be altered other than by either changing the basic task or changing knowledge level of the learner. They argue that it can only be altered by changing the nature of what is learned or by the act of learning itself.

Extraneous Load

Working memory load is not only imposed by the intrinsic complexity of the material that needs to be learned. Extraneous cognitive load is imposed by instructional procedures to the process of learning that are less than optimal. Unnecessary detail, insufficient instruction, inappropriate orders of delivery and poor use of resources can all contribute to extraneous cognitive load (Van Merrienboer & Sweller, 2005). Beckmann (2010) suggests that element interactivity is a major source of working memory load underlying extraneous as well as intrinsic cognitive load. *"If element interactivity can be reduced without altering what is learned, the load is extraneous; if element interactivity only can be altered by altering what is learned, the load is intrinsic"* (Beckmann, 2010). In educational setting, especially in the context of third level education, instruction material is often presented in split-attention format, which increases extraneous cognitive load compared to physically integrated formats (Sweller, Chandler, Tierney, & Cooper, 1990). For example, there is a split-attention effect, Chandler and Sweller (1992), which occurs when cross-referencing sources from different places. Van Merrienboer and Sweller (2005) have shown that tasks such as reading from slides or looking up data tables while reading textual information also splits attention, known as the modality effect. They suggest that such effects can be excluded from learning by instructional interventions such as providing materials with all required information in one place.

Germane Load

Germane load is the cognitive load on working memory generated by explicit instructional designs aligned to task difficulty. It is created by activities especially designed to create scheme construction. (Sweller, Van Merrienboer, & Paas, 1998). Worked examples at the appropriate stages of learning impose germane load. Germane cognitive load also can be specified in terms of element interactivity. In contrast to the emphasis by intrinsic and extraneous cognitive load on the characteristics of the material, germane cognitive load is concerned only with learner's characteristics. It refers to the working memory resources that the learner devotes to cope with the intrinsic cognitive load (Sweller, 2010). Germane load does not constitute an independent source of cognitive load.

Sweller and colleagues, with their attempt to define cognitive load within the discipline of Educational Psychology and for instructional design, believed that the three types of cognitive load are additive. This implies that the total cognitive load experienced by a learner is the sum of all three types of cognitive load.

2.2 Instructional Design

Cognitive load theory has used the human cognitive architecture to devise cognitively effective and efficient instructional procedures (P. A. Kirschner, Sweller, Kirschner, & Zambrano R., 2018). Instructional design is a field of study that attempts to combine education, psychology and communication to produce the most effective ways or strategies for a specific group of learners. In other words, it is the design of instructional materials. The principles of instructional design also considers how participants learn, what medium of delivering the topic will be most effective and meaningful so that they can better understand the topics being taught.[1] Sweller (2011) acknowledge that while cognitive load theory is not unique in using human cognition to generate instructional procedures, it is regrettably rare for instructional design to be based on human cognitive architecture.

2.2.1 Direct Instructions

The premise for acquiring knowledge in CLT is that learners have to be instructed by means of direct instructional designs (Sweller et al., 2019; Sweller, 2016). It assumes that working memory can only process explicit and direct instructions. Direct

[1] Purdue University. *What is instructional design?* Retrieved from https://online.purdue.edu/blog/education/what-is-instructional-design

instruction has come to have many different meanings, all of which are associated with some form of structured teaching. Direct instructional guidance is defined as providing information that fully explains the concepts and procedures that students are required to learn as well as learning strategy support that is compatible with human cognitive architecture (P. Kirschner, Sweller, & Clark, 2006). It is a systematic attempt to build effective academic instruction that includes all of the school-based components necessary to produce academic growth (Slocum, 2004). To teach effectively and efficiently, big ideas must be conveyed to the learners in a way that clear, simple, and direct. According to Peterson (1979), direct instruction has the following characteristics: academic focus, little learner choice of activity, instructor-centered focus, potential to cater for a large group, use of factual information and controlled sessions.

According to the National Institute for Direct Instruction, direct instructions operate on some key principles[2], some of which are:

- All students can be taught.
- All details of instruction must be controlled to minimize the chance of students' misinterpreting the information being taught and to maximize the reinforcing effect of instruction.

2.2.2 Community of Enquiry Technique

Inquiry is proposed as teaching and learning technique that is deeply linked with a continuous auto-corrective process of knowledge development. Through the process of inquiry, an unsatisfactory situation can be converted into satisfactory by connecting all of its constituent into a coherent and unified whole (Dewey, 2007). Inquiry techniques are proposed in the educational context to improve the comprehension of complex learning tasks (Garrison, 2007). The community of inquiry may be defined

[2] Engelmann, S. (n.d.).Basic philosophy of direct instruction (DI). Retrieved from https://www.nifdi.org/what-is-di/basic-philosophy

as ‘a teaching and learning technique, an instructional technique of a group of learners who, through the use of dialogue, examine the conceptual boundary of a problematic concept proceeding all the parts this problem is composed of in order to solve it’ (Orru, Gobbo, O’Sullivan, & Longo, 2018). The collective work of the group creates what is known as a collective working memory effect which enables learners to share the working memories among multiple participants that share the same task. The assumption behind this effect is that the use of working memory of many individuals can reduce the overall cognitive cost of the task at hand. This also implies that the working memory capacity of the group may be increased (P. A. Kirschner et al., 2018).

A theory that is linked to the inquiry method is the Social Constructivism Theory, which is a variable of the Theory of Constructivism. Since learning is considered to be an active process with the learners constructing their knowledge based on experience and reflecting on this experience, social constructivism focuses on the social and cultural context which shapes the construction of knowledge (Simina, 2012). Social constructivism upholds the idea that human development is socially situated and knowledge is constructed through interaction with others. Social constructivist methods are implemented in educational institutions, including third level educational institutions. Instructors encourage the students to actively participate in ongoing discussions and use the responses from the students to create lesson plans and modify content where necessary. this allows opportunities for students to create associations for the subject topic.

According to Sweller (2009), constructivism ignores the human cognitive architecture. As a consequence, constructivism cannot lead to instructional design aligned to the way humans learn, so they are set to fail due to lack of explicit instructional designs (P. Kirschner et al., 2006). However, this idea is argued stating that in educational psychology, there is no relationship between inquiry methods and direct instruction (Jonassen, 2009). This is based on the premise that the direct instructions approach and the inquiry approach come from different theoretical assumptions and they utilise different methods. In order for the two methodologies to be compared, they have to share a learning outcome or same dependent variable.

The choice of instructional design / approach should depend on the educational objective an instructor wishes to attain(Peterson, 1979). If the learning outcome is the improvement of inquiry skills, then direct instructions should not be used. However, if the outcome is to teach basic understanding of topics, then direct instructions would be more appropriate. Another consideration would be the type of student who is being taught. A student with low abilities might need the structured delivery of the direct instructions, whereas a student with high abilities may benefit from the social inquiry with others in a less direct approach. The effectiveness of the methods also requires an element of decision making from the instructor.

2.3 Mental Workload

Despite all the years of research, no proven measure of the three cognitive loads have emerged, based on empirical research. This has lead to the development of many models by employing many techniques (Longo, 2011, 2012, 2014; L. M. Rizzo & Longo, 2017). The concept of cognitive load is mainly employed in the education field, whereas the concept Mental Workload, a psychological construct strictly connected to cognitive load, is employed is mainly employed in ergonomics (Longo & Leva, 2017). The former relates to working memory resources only, whereas latter takes into account other factors as the level of motivation, stress and the physical demand experienced by participants as a consequence of the task. Despite of their different fields of research, they both assume that working memory limits must be considered to predict performance while accomplishing an underlying task. Although the field of educational psychology is struggling to find ways of measuring mental workload of learning tasks, there is an entire field within Ergonomics devoted to the design, development and validation of reliable measures of mental workload. Mental Workload (MWL) is defined as the volume of cognitive work necessary for an individual to accomplish a task over time

(Longo, 2015; L. Rizzo, Dondio, Delany, & Longo, 2016).

The measurement of cognitive load is of crucial importance for instructional research. The few efforts in instructional research to measure cognitive load are almost exclusively concerned with performance measures (Paas, Van Merrienboer, & Adam, 1994). Different techniques, with different advantages and disadvantages, have been proposed in education to measure mental workload (cognitive load) and they can be clustered in two main groups: Subjective and Objective measures (Plass, Moreno, & Brünken, 2010).

Subjective Measures are more suitable to be applied in an educational context and in general are easy to administer and analyse, in contrast to objective measures. Subjective Measures, also referred to as self-reported measures, rely on the individual's perceived experience of the interaction with a learning task. It is based on the assumption that only the individual involved in the task can provide an accurate and precise judgement about the experienced mental workload, as employed in a number of studies (Junior, Debruyne, Longo, & O'Sullivan, 2019; Moustafa & Longo, 2019). The perception of the individual can be gathered through means of a survey or questionnaire. Subjective measures include both Uni-dimensional approaches and multidimensional approaches, which have been conceptualised, applied and validated. The most commonly used subjective measures are uni-dimensional. They provide an index of overall workload, but provide no information about its temporal variation. Multidimensional measures can determine the source of mental workload. They seem to be the most appropriate types of measurement for assessing mental workload because they have demonstrated high levels of sensitivity and diagnosticity (Rubio Valdehita, Ramiro, García, & Puente, 2004).

2.3.1 Uni-dimensional Measure

Paas (1992) equals the effort of learners to the overall cognitive load, thus mental effort alone , one variable, can measure the different types of load (Paas, 1992). The modified

Rating Scale of Mental Effort (RSME) (F. Zijlstra & Doorn, 1985) is a unidimensional instrument used to measure subjective mental workload. This assessment procedure is built upon the notion of effort exerted by a human over a task. A subjective rating is required by an individual through an indication on a continuous line, within the interval 0 to 150 with ticks each 10 units, each accompanied by a descriptive label indicating a degree of effort (F. R. H. Zijlstra, 1993). Example of labels are 'absolutely no effort', 'considerable effort' and 'extreme effort'. The overall mental workload of an individual coincides to the experienced exerted effort indicated on the line. RSME requires no special device to record the measurements. The method is simple, and it allows for quick response and applicability without interfering with the work of the individuals (Ghanbary Sartang, Ashnagar, & Sadeghi, 2016). RSME has shown a poor diagnostic power, nevertheless it has demonstrated a good degree of sensitivity across different empirical studies (F. R. H. Zijlstra, 1993).

2.3.2 Multidimensional Measure

A well-known multidimensional subjective measure is the NASA Task Load Index (NASA-TLX). Although widely employed in Ergonomics, this has been rarely adopted in Education. (De Jong, 2010)(2010) argues that the use of this multidimensional measure is exceptional in education. A few studies have confirmed its validity and sensitivity when applied to educational context (Fischer, Lowe, & Schwan, 2008; Gerjets, Scheiter, & Catrambone, 2006; Kester, Lehnen, Van Gerven, & Kirschner, 2006). It focuses on six different components of load. These represent independent clusters of variables: mental, physical, and temporal demands, frustration, effort, and performance (Hart & Staveland, 1988). In general, the NASA-TLX has been used to predict critical levels of mental workload that can significantly influence the execution of an underlying task.

To collect ratings for the dimensions, a twenty grade scale is utilised for each dimension. A score from 0 to 100, at intervals of 5, is collected on each scale from the respondents. Then a weighting procedure is used to connect all the individual ratings together. A paired comparison task is required from the respondent before the workload calculation can be undertaken. These paired comparison allows to choose the more pertinent dimension over all pairs of the six dimensions. A workload score from 0 to 100 is calculated for each task by multiplying the weight by the individual dimension scale score, summing aacross scales, and then dividing by 15 (the total number of paired comparisons). The formula to calculate NASA-TLX score is as follows:

$$NASA : \big[0..100\big] \epsilon \ \Re \qquad NASA = \frac{\big(\sum_{n=1}^{6} d_i \times w_i\big)}{15}$$

where d$_i$ is the score of each question while w$_i$ is the weight for that question generated by the pairwise comparison procedure. There is a modified version the NASA-TLX, the raw NASA-TLX (RAWNASA). With this technique, the weighting process is eliminated. (Hart, 2006)(2006) summarised the results of research conducted along various studies in which the RAWNASA method was compared with the NASA-TLX. Based on those studies, RAWNASA was found to be mode sensitive (Hendy, Hamilton, & Landry, 1993), less sensitive (Liu & Wickens, 1994) or equally sensitive (Bittner Jr., Byers, Hill, Zaklad, & Christ, 1989). NASA-TLX, RAWNASA and RSME are reliable and valid measures of mental workload when applied in the educational context (Longo & Orru, 2019).

2.4 Instructional Efficiency

Efficiency of instructional designs in education is a measurable concept (Longo & Orrú, 2020). Efficiency in the context of problem-solving, learning and instruction is the capacity to achieve established goals at the minimal expense of resources (Hoffman & Schraw, 2010). Paas et al. (1993; 1994) suggest that combining performance and mental effort measures allow the calculation of an index of mental efficiencies. Studies that investigated processing instructional efficiency made use of uni-variate scores to compare the impact of an experimental condition on a control group. Sweller(2010) argues that instructional effectiveness will be compromised by the extent that instructional choices require learners to devote working memory resources to dealing with elements imposed by extraneous cognitive load. They also state that, at a basic level, understanding efficiency is an essential precursor to assessing educational effectiveness and improvement. Various studies that propose measures of efficiency have been conducted in the past. The most common and widely used measures / models of efficiency will be discussed below.

2.4.1 Models of Instructional Efficiency

Deviational model

In search of a single measure to determine the relative efficiency of instructional conditions in terms of learning outcomes, Paas and Van Merrienboer (1993) developed a computational approach for combining measure of performance with measure of mental effort to attain efficiency. This was characterised as the Instructional Condition Efficiency. This is referred to as the **Deviational model** of efficiency by Hoffman and Schraw (2010) because this model computes the difference between a standardised score of performance and a standardised score of effort. The reasoning behind this formula is based on the assumption that the resulting efficiency is high when an individual experiences high performance and low effort. Conversely, the resulting efficiency is low when an individual experiences low performance and high effort (Paas & Van Merrienboer, 1993). The deviational model of efficiency computes a measure of efficiency based on how the participant performs relative to the group (Hoffman & Schraw, 2010). It measures the distance from the observed score to the ideal efficiency slope. The deviational model provides a group-referenced score representing an individual efficiency that requires scores to be converted to a common scale. Efficiency score using the deviational model is computed using the following formula:

$$Efficiency = \frac{(ZP - ZR)}{\sqrt{2}}$$

where ZP = Standardised Performance Score and ZR = Standardised Effort Score.

If ZP - ZR > 0, then efficiency is positive. If ZP - ZR < 0, then efficiency is negative. According to the authors, the highest efficiency condition occurs when performance was maximum and effort was minimum. The lowest efficiency corresponds to the lowest performance and highest effort(Paas & Van Merrienboer, 1993). There are concerns expressed by Hoffman and Schraw(2010) that the efficiency score computed by the deviational model is problematic because the standardised scores are affected by variability and performance of others within the group. They also expressed that the results should be interpreted cautiously although the results may be mathematically identical in magnitude and direction, as they may be conceptually incommensurate.

The original formula of calculating instructional efficiency proposed by Paas and Van Merrienboer (1993) was based on performance and mental effort invested to attain this performance. Subsequently, further studies by other researchers, for example (Tindall-Ford, Chandler, & Sweller, 1997), have combined mental effort spent during training with performance to calculate the instructional efficiency. The original approach reflects learning efficiency, the latter approach is argued to reflect both training and learning efficiencies (Tuovinen & Paas, 2004).

Likelihood model

One of the measures of efficiency developed within the education context is based upon the likelihood model put forward by Hoffman and Schraw(2010). Efficiency is this model is computed as a ratio of work output to input. In other words, a ratio of performance to perceived mental effort. Output is identified with learning and input is identified with time, work or effort (Smith & Street, 2005).

Efficiency score using the likelihood model is computed using the following formula:

$$Efficiency = \frac{P}{R}$$

where P = Raw score of performance and R = Raw score of perceived effort.

An estimation of the rate of change of performance is calculated by dividing P by R and the resulting ratio represents the individual efficiency based on individual scores of performance and effort (Hoffman & Schraw, 2010). The ratio ranges from zero to extensive positive values; it goes towards zero when performance is low and effort is high (low efficiency) and conversely, goes towards the extensive positive value when performance is high and effort is low (high efficiency). The authors argue that, compared to the deviational model of efficiency, the likelihood model provides an unambiguous measure because the inputs are not standardised scores, and there is no restrictions in the range of efficiency scores. However, the resulting efficiency here is always going to be positive. it must be interpreted with caution because the formula assumes that the work input is not zero (Hoffman, 2012). It is also acknowledged that efficiency scores based on this model is supposedly more reliable and sensitive to minor effect size changes compared to the deviational model. An extension on this likelihood has been employed by Kalyuga and Sweller(2005) where an extra reference to a *critical value* is used, under or above which the efficiency can be considered negative or positive(Kalyuga & Sweller, 2005). The authors suggest to obtain the critical value by dividing the maximum performance score by maximum effort exerted by a learner in order to establish whether that learner is competent or not. The ratio of the critical is based on the underlying assumption that an instructional design is inefficient if a learner invests maximum effort in a task without reaching maximum performance and vice-versa. (Kalyuga & Sweller, 2005). Through this extended formula, the model evolves from one being able to define only positive efficiency score to one capable to defining a positive / negative efficiency.

Multidimensional model

Tuovinen and Paas(2004) extended the original deviational model formula proposed by Paas and Van Merrienboer(1993) and the adapted deviational models of other researchers by including a third dimension to the model. The authors referred to this as the "3D Instructional condition efficiency model". This model was developed on the assumption that it is feasible for two individuals to achieve the same performance score, while experiencing different levels of mental effort. It is assumed that the individual who experienced the least effort, while achieving the same performance was able to learn the topic more efficiently compared to the other individual. This three factor or dimension approach utilises a performance component as well as effort expended during both learning and test conditions. The authors claim that the proposed model should be more sensitive to the individual's learning than the performance score alone and therefore, provides a better efficiency. Efficiency score using the model is computed using the following formula:

$$Efficiency = \frac{(ZP - ZR_L - ZR_T)}{\sqrt{3}}$$

where ZP = Standardised Performance Score, ZR$_L$ = Standardised Learning Effort Score and ZR$_T$ = Standardised Test Effort Score

In a comparison study conducted by Hoffman(2012), there exists a computational difference in the formula originally proposed by Tuovinen and Paas(2004). In the comparative study, the formula is given as follows:

$$Efficiency = \frac{(ZP - ZR_L + ZR_T)}{\sqrt{3}}$$

According to the author, the 3D Instructional condition efficiency model uses subtraction to calculate a difference. This formula relies upon non-associative mathematical properties, meaning that the inverse of mathematical operations will not produce the same difference or quotient (Hoffman, 2012). However, there seems to be no evidence to support this explanation contained within the research document. Efficiency scores that rely on the standardised scores of performance and perceived effort are most useful determining the magnitude of difference.

Parabolic model

Johnes, Silva, and Thanassoulis (2017) state that high efficiency occurs when outputs from education (such as test results) are produced at the lowest level of financial, cognitive or temporal resources (Johnes et al., 2017). The general assumption is that as the difficulty of a task increases, so does the effort required to complete it; and as a result, the performance usually decreases. Excessive workload caused by a task using the same resource can create problems and result in errors or lower task performance. When workload increases it does not mean that performance always decreases: performance can also be affected by workload being too high or too low (Nachreiner, 1995). A high level of mental workload can be related with a high level of focus on the task whereas a low level might means little or no mental resource allocated to a task. Since the working memory is limited in its capacity, it is important not to exceed its limits in order to get the best performance. An optimal level of mental workload facilitates the learning process, whereas a high level (overload) or a low level (underload), hampers the learning phase (Longo, 2016). Motivated by these statements, and the general assumption of what constitutes high / low efficiency, this novel model of instructional efficiency was developed by Dr. Luca Longo, Technological University Dublin.

The underlying principle behind the parabolic model of efficiency is an assumption based on the expected relationship between performance and mental workload. This is represented as a parabola (curved black line) in Figure 2.1. The expectation with this parabola is that for an amount of exerted mental workload, a certain performance should be achieved and vice-versa; to achieve specific performance, a certain amount of mental workload should be exerted. Another important assumption for this model is that the individual is expected to have no prior knowledge on the activity / topic. According to the parabola, the Maximum Efficiency is expected to be achieved when the mental workload exerted is at 50% of maximum capacity (MWL_{max} / 2) and the performance is at maximum or 100% (P_{max}). This is referred to as the **ideal point**. Additionally, the point on the plane is where a person achieves zero performance and exerts zero mental workload, zero efficiency, is referred to as the **worst point**.The way that the mental workload should be measured for this particular model is not specified, however it is expected that any measure of mental workload could be used, as long as it is defined clearly.

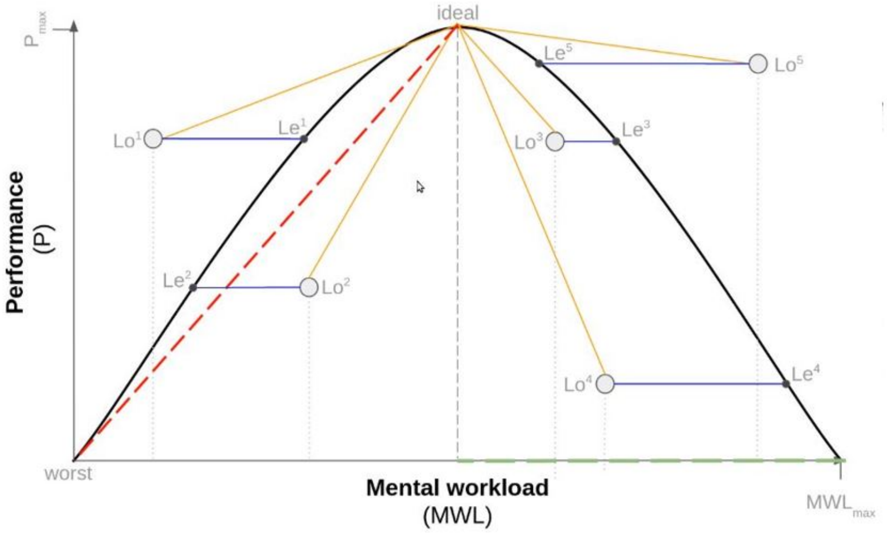

Figure 2.1: Graphical representation of the parabolic model of efficiency proposed by Dr. Luca Longo, Technological University Dublin

Four points are represented over a two-dimensional Cartesian coordinate system.; ideal point, worst point, the expected point and the observed point. They are as follows:

Ideal	$((MWL_{max} / 2) , P_{max})$
Worst	(MWL_0 , P_0)
Expected	(Expected MWL , Expected P) [on the parabola]
Observed	(Actual MWL , Actual P)

where MWL = Mental Workload and P = Performance.

The parabolic model calculates an efficiency score on the premise of distance between the various points as specified above. Efficiency score using the parabolic model is computed using the following formula:

$$Efficiency = \frac{\left[1 - \frac{D(Lo^x,ideal)}{D(worst,ideal)}\right] + \left[1 - \frac{\left|D(Lo^x,Le^x)\right|}{MWL_{max}/2}\right]}{2}$$

where **D** = a measure of distance between two specified points, $\mathbf{MWL}_{max}$ = Maximum rating of Mental workload, **o** = observed point, **e** = expected point, **ideal** = ideal point, **worst** = worst point. There are two other elements to note from this formula which are provided for reference; **L** = The reference learner and ^{x} = nth observation. These can be observed in Figure 2.1 for better understanding.

The idea behind this model of efficiency is very theoretical and much more complex than meets the eye and certainly more empirical research is required to prove the validity and sensitivity of this model (Longo, 2018). According to CLT, the parabolic model could be potential indicator of Germane load. For example, an individual who achieves zero performance after exerting maximum workload would receive a very low efficiency score, but not zero. This is because the individual would be engaged in the activity to the point where there is an overload of mental workload but failed to achieve any performance. However, the activity / instruction delivery should not be penalised for such an outcome. The instruction delivery would still be considered somewhat efficient because there is active participation from the individual. There is no concept of a negative efficiency with this model, similar to the likelihood model, because it is not a relative scale. The range of values for efficiency score based on this model of efficiency is between 0 and 1.

2.5 Gaps and Research Question

It is vital to develop models of efficiency that are relevant to education and a wide range of other disciplines. Research in education and psychology currently relies on competing models, despite the fact that little attention has been paid to differences among these models and the implications of these differences for understanding and improving efficiency (Hoffman & Schraw, 2010).

Formulas of instructional efficiency exist for the evaluation of instructional conditions, based upon combinations of performance and perceived mental effort / workload. However, the gap that emerged from the literature review points to the lack of comparison between the different models of efficiency in the third-level educational domain. Moreover, a lack of literature on the parabolic model exists. The parabolic model is not a proven model yet, it is only a theoretical concept. Therefore, to be recognised and gain credibility, the model needs to be examined, evaluated and compared with other state-of-the-art models to determine it's validity and sensitivity. All the models are theoretically different to each other and should not behave like each other. Therefore, the assumption here is that there would be moderate correlation between them. A moderate correlation is expected to ascertain that the different models of efficiency are measuring the same conceptual outcome.

Research Question

The research question being proposed in this study is as follows:

To what extent can we compare and discriminate between the control and experimental groups using the parabolic model when compared to the other models of instructional efficiency in third-level classes?

Chapter 3

Design and Methodology

This chapter explains the design of the experimental framework with the aim to solve the empirical research question.

3.1 Research Hypothesis

Given the research question in chapter 2, a primary research experiment was designed, and the following research hypothesis was set:

"H_1: IF the Parabolic model of efficiency (PM) is employed to compute teaching and learning instructional efficiencies in 3^{rd} level classrooms, THEN it is expected that it exhibits higher Sensitivity ***AND*** *higher Discriminant Validity than the Likelihood (LM) and Deviational (DM) models* ***AND*** *moderate to strong Concurrent Validity with them."*

The implementation of the experiment will take place in several stages. The first stage is data understanding which includes data gathering. The second stage consists of data preparation to proceed with the study. The third stage consists of data modelling which describes the different models of instructional efficiency employed and how the efficiencies will be calculated. The final stage consists of model evaluation which explains the various ways in which the models will be evaluated. Figure 3.1

illustrates the flow of the experiment.

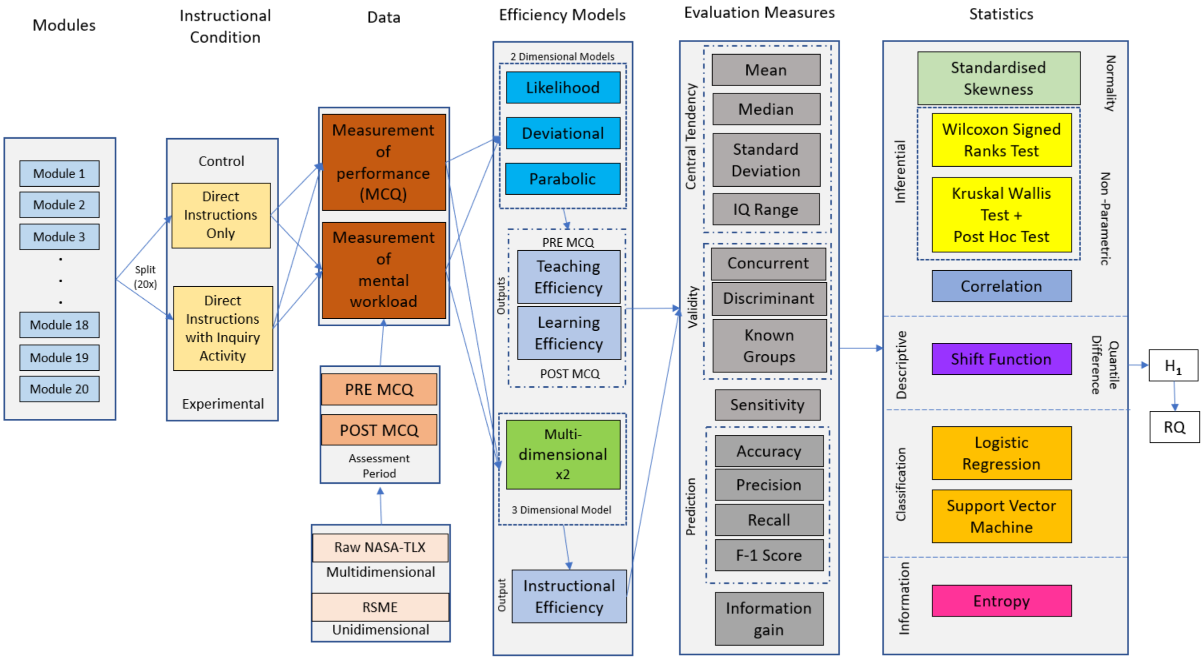

Figure 3.1: Flow of Experiment Design

3.2 Data Understanding

Data for this experiment was collected by Giuliano Orru from various classes and modules in Technological University Dublin (Orru et al., 2018; Orru & Longo, 2019a, 2019b, 2020). The participants associated to the modules were informed of the criteria for the voluntary participation in the experiment and complete anonymity of any published data. Study Information, along with participant form were distributed to each participant at the beginning. The consent form was approved by the Ethical Committee of Technological University Dublin. After the participants completed the study information and consent forms, the participants were divided into two groups at random: control and experimental. The participants in each class were divided evenly as far as practicable.

The experiment compares two instructional design conditions. The first design followed the direct instruction approach to learning, while the second design extended that with a collaborative inquiry activity designed to replicate the community of inquiry approach to learning. The former approach involved a theoretical explanation of a chosen topic, whereby the instructor presented the information through direct instructions. The direct instructions were specific and clear, aimed at facilitating the learning and problem solving. The latter approach extended the instructions with a guided inquiry activity amongst participants based on some cognitive trigger questions. Both groups received direct instructions, while only the experimental group subsequently participated in the collaborative inquiry activity. The purpose of this design is to establish whether the extension improves the efficiency of learners compared with learners who receive direct instructions only.

After the topic was presented to the class by the instructor, the control group participants received questionnaires aimed at quantifying the effort and mental workload they experienced, using Rating Scale Mental Effort (RSME) and the NASA task load index (NASA-TLX) factors respectively, along with a multiple-choice questionnaire (MCQ) associated to the topic taught. The experimental group was split into teams of three or four participants for the inquiry activity. The participants discussed and exchanged information related to the topic and formed informed agreements collaboratively. The participants then wrote the shared answers individually to the cognitive trigger questions. After the activity, the experimental group participants received questionnaires aimed at quantifying the effort and mental workload they experienced, along with a multiple-choice questionnaire (MCQ) associated to the topic taught, similar to the control group. Once the participants in both groups completed the MCQ, they were provided with a further questionnaire aimed at quantifying the effort and mental workload they experienced after completing the MCQ. Filling the questionnaire on both occasions allows the researcher to compute both the training efficiency and the learning efficiency, as they are related to different stages of the learning process.

3.3 Data Preparation

The dataset will be inspected for missing values and any cases with missing values will be considered for removal. The possibility of removing a variable will also be considered for any with consistent missing values across a range of cases, if the variable is not considered important for any of the efficiency models employed.

Outliers and any possible anomalies will be detected. Removal of outliers will be considered as they could potentially influence the outcome. Removal of outliers will also depend on the variable under consideration.

Standardised PRE-MCQ effort scores, POST-MCQ effort scores and MCQ scores will be computed for use with some of the models of efficiency and will be added to the dataset as separate variables.

The raw NASA task load index (RAWNASA) scores will be computed for each case, both before the MCQ (PRE-MCQ) and after the MCQ (POST-MCQ) using the answers provided by the participants in the questionnaire. The 6 individual ratings of NASA-TLX variables will be transformed into an overall combined score for RAWNASA. There is no calculations necessary for RSME scores. It is a unidimensional, subjective rating indicated by the participant on a continuous line, with the interval of 0 to 150 with ticks every 10 units. It is simple and sensitive.

3.4 Modelling

The principal aim of this stage is to create different models of instructional efficiency widely used in instruction & education and compute efficiency scores using these models to determine which model best discriminates between the control and experimental groups, as well as compute the instructional efficiency more accurately.

Five models of instructional efficiency will be modelled to compute the efficiency scores for each observation. The models are as follows:

1. Likelihood model of efficiency
2. Parabolic model of efficiency
3. Deviational model of efficiency
4. Multidimensional model of efficiency (Original)
5. Multidimensional model of efficiency (Modified)

Both training and learning efficiency scores will be computed for the following models of efficiency: Likelihood model, Parabolic model and Deviational model. Since the multidimensional model utilises both Learning and Test efforts to compute the efficiency, only Instructional Efficiency, will be computed. An instructional efficiency score will be computed for each variant of multidimensional model, resulting in two instructional efficiency scores.

Since two measures of effort / mental workload is being considered in this experiment, efficiency scores will be calculated using both RSME and RAWNASA for each model of efficiency, resulting in a total of 16 different efficiency scores. They are as follows:

- Likelihood model of efficiency
  - Training efficiency with RAWNASA (TR.EFF_LM_RAWNASA)
  - Learning efficiency with RAWNASA (LR.EFF_LM_RAWNASA)
  - Training efficiency with RSME (TR.EFF_LM_RSME)
  - Learning efficiency with RSME (LR.EFF_LM_RSME)

- Parabolic model of efficiency
  - Training Efficiency with RAWNASA (TR.EFF_PM_RAWNASA)
  - Learning Efficiency with RAWNASA (LR.EFF_PM_RAWNASA)
  - Training Efficiency with RSME (TR.EFF_PM_RSME)
  - Learning Efficiency with RSME (LR.EFF_PM_RSME)
- Deviational model of efficiency
  - Training efficiency with RAWNASA (TR.EFF_DM_RAWNASA)
  - Learning efficiency with RAWNASA (LR.EFF_DM_RAWNASA)
  - Training efficiency with RSME (TR.EFF_DM_RSME)
  - Learning efficiency with RSME (LR.EFF_DM_RSME)
- Multidimensional model of efficiency (Original)
  - Instructional efficiency with RAWNASA (INS.EFF_3DM_RAWNASA)
  - Instructional efficiency with RSME (INS.EFF_3DM_RSME)
- Multidimensional model of efficiency (Modified)
  - Instructional efficiency with RAWNASA (INS.EFF_3DM_RAWNASA2)
  - Instructional efficiency with RSME (INS.EFF_3DM_RSME2)

Once the efficiency scores are computed, they will be added to the dataset as separate variables.

The dataset will be explored and inspected for normality and skewness. To determine the normality of the distribution, a uni-variate analysis will be performed. The analysis will include graphical representations and skewness tests. The variable will be considered as normally distributed if the standardised score of skewness is between +/- 2 (George & Mallery, 2010).The Shapiro Wilk test will not be used to check for normality as the test is sensitive and has a bias by sample size of the dataset. [1] Small sample sizes result in low statistical power for normality tests. This means that substantial deviations from normality will not result in statistical significance. Normality tests are only needed for small sample sizes, but this is also the situation in which they perform poorly.

Validity is an important aspect of effective research. Validity is the extent to which an instrument measures what it is meant to measure (Krabbe, 2017). To assess the validity of the different Efficiency Scores, two sub-forms were selected, namely Concurrent and Discriminant.

Concurrent validity is a type of criterion-related validity which endeavours to relate results of one particular instrument to another external criterion. Concurrent validity can be demonstrated, if the efficiency scores from one model correlates highly with the efficiency score from another model. The advantage of concurrent validity is that concurrent validity between two instruments can be demonstrated simultaneously (L. Cohen, Manion, & Morrison, 2005). Concurrent validity will be assessed by performing a correlation test between the training efficiency scores of all three 2-dimensional models (LM, PM and DM) in pairs and the learning efficiency scores of three 2-dimensional models in pairs. Concurrent validity will be tested for both RSME and RAWNASA measures. Both parametric (Pearson) and non-parametric (Spearman) tests will be considered based on the normality of the distributions and other assumptions of the tests.

Discriminant validity validates the degree to which the two scores of efficiency, expected to be theoretically unrelated, are in fact unrelated (Carlson & Herdman, 2012). Discriminant validity will be assessed by performing a correlation test between the training and learning efficiency scores for all three 2-dimensional models. Dis-

[1] Stephanie Glen. ”Shapiro-Wilk Test: What it is and How to Run it” from https://www.statisticshowto.com/shapiro-wilk-test/

criminant validity will be tested for both RSME and RAWNASA. Both parametric (Pearson) and non-parametric (Spearman) tests will be considered based on the normality of the distributions and other assumptions of the tests.

Sensitivity is the extent to which the efficiency scores can detect changes in the instructional design and discriminate between the groups. Sensitivity will be assessed by checking whether the distributions of all efficiency scores are statistically significant different across the modules and the groups. Sensitivity will be assessed in three ways; Known Groups validity, Shift function and Classification.

Known Groups validity, also known as extreme-groups validity, is a strategy that indirectly assesses the validity of a set of observations by demonstrating that the set's output varies systematically depending upon known performances of the construct that the scale is intended to measure (Virues-Ortega, Montaño-Fidalgo, Froján-Parga, & Calero-Elvira, 2011). Known Groups validity can be demonstrated when a test can discriminate between the control and experimental group which are known to differ on the basis of the instructional design. Known Groups validity will be assessed by checking whether the distributions of all efficiency scores are statistically significant different across the modules. ANOVA and their non-parametric equivalent, Kruskal-Wallis will be considered based on the normality of the distributions and assumptions of the tests, as well as the relevant Post-Hoc tests.

In addition to the Known Groups validity, which analyses differences between distributions based on a central tendency measure, e.g., the median, a shift function will also be employed in the experiment. The general assumption when comparing two distributions is that they differ only in central tendency, not in other aspects. This consideration is not robust as there is no reason a priori to assume this. Effects can occur in the tails of the distributions too. To account for this, the entire distribution needs to be compared. The shift function plots the differences between quantiles of two different groups as a function of one group. This is used to visualise the comparison between two groups and determine how, and by how much, two distributions differ. The shift function will be employed on all models to identify which model differentiates the differences between the groups better. The Shift Function will be employed for all 16 efficiency scores across all modules.

In an effort to understand which efficiency score provides the most information about whether the observation belongs to the control or experimental group, Entropy and Information gain will be calculated on the dataset using the 16 scores of efficiency. Entropy is a measurement of uncertainty in the data (Murphy, 2012). It provides a measure of purity and quantifies how much information there is in a random variable, or more specifically its probability distribution. Entropy of a dataset can be viewed in terms of the probability distribution of observations in the dataset belonging to one class or another. In this case, the probability distribution of observations in the dataset belonging to the control group v experimental group. In the context of classification, entropy measures the diversification of the class labels [2]. Entropy for the ”group” variable will be calculated over the complete, unsegmented dataset, as well as over data sets segmented by modules.

Information gain, or I.G., is a measure of reduction in Entropy by transforming the dataset in some way. Information gain is calculated by comparing the Entropy of the dataset before and after a transformation. It measures how much ”information” a variable provides about the class. It is commonly used in training a decision tree by evaluating the information gain for each variable, and selecting the variable that maximizes the information gain, which in turn minimizes the entropy and best splits the dataset into groups for effective classification. [3]. Information gain will be calculated for each efficiency score variable over the complete, unsegmented dataset, as well as over data sets segmented by modules.

[2] Information Gain, Gain Ratio and Gini Index (Phung, 2020)

[3] Information Gain and Mutual Information for Machine Learning (Brownlee, 2019)

To determine the discriminating capacity of the different models of efficiency, different classification models, or classifiers, will be built for each efficiency score variable within the dataset. This will be split into two classification problems:

1. Classify the group
2. Classify the module

For each classification problem, a classifier will be built for each efficiency score, where the efficiency score will be the predictor variable and the group or module will the target variable.

The classification models will be built using two different learning approaches:

1. Logistic regression
2. Support vector machine

Logistic Regression (LR) is a powerful statistical way of modelling qualitative outcome with one or more predictor variables. It measure the relationship between the target variable and the predictors by estimating probabilities using a logistic function. To predict the group, a binomial logistic regression model will be employed because there are 2 categories in the group variable. To predict the modules, a multinomial logistic regression model will be employed because there are 20 categories in the module variable. A total of 16 binomial regression models and 16 multinomial regression models will be built.

Support vector machines (SVM) are supervised learning models that partition a feature space into two or more groups. This is achieved by finding an optimal means of separating the groups based on the known class labels. Support vector machines apply a simple linear method to the data but in a high-dimensional feature space non-linearly related to the input space (Karatzoglou, Meyer, & Hornik, 2006). Support vector machines are capable of carrying out non-linear partitioning by means of the kernel function which transform the data in order to accommodate a non-linear boundary between the classes. The SVM classifiers will be modelled separately with 4 different types of kernels: Linear, Radial, Polynomial and Sigmoid. The SVM models will be used to predict both group and module using each type of kernel. A total of 128 SVM models will be built, 32 per kernel type.

Once the above modelling steps are carried out, the various models of instructional efficiency will be evaluated as outlined in Chapter 3.5.

3.5 Model Evaluation

The central tendency of all the numeric variables will be examined. Central tendency measures include, but not limited, the following:

- Mean (M)
- Median (Mdn)
- Standard Deviation (SD)
- Inter-Quartile Range (IQR)

A significance level α of 0.05 will be adopted for this research. If the p-value is < 0.05, then it will be deemed statistically significant.

Cohen's heuristics on effect size will be adopted for all relevant statistical tests. Cohen suggests to employ the following rule of thumb for interpreting results related to effect size of correlation (J. Cohen, 1988; J. Cohen, Cohen, West, & Aiken, 2003):

- $x < 0.1$ = neutral correlation
- $0.1 \leq x \leq 0.3$ = small/weak correlation
- $0.3 \leq x \leq 0.5$ = medium/moderate correlation
- $x > 0.5$ = large/strong correlation

Concurrent validity will be tested as described in Chapter 3.4 and will be evaluated. Concurrent validity will be demonstrated by the resulting spearman's rho (r_s). The correlations between PM and other models (LM & DM) should be statistically significant and have an average $r_s \geq 0.3$.

Discriminant validity will be tested as described in Chapter 3.4 and will be evaluated. Discriminant validity will be demonstrated by the resulting spearman's rho. The correlations between parabolic model's efficiency score pairs should be statistically significant and have an average r_s lower than the other models' correlation pairs.

Known Groups validity will be tested as described in Chapter 3.4 and will be evaluated. Known Groups validity will be demonstrated if the efficiency discriminates between the control and experimental groups. The efficiency score which discriminates between the two groups across all modules with the most statistically significant differences will be deemed the model with better sensitivity.

Shift Function will be employed as described in Chapter 3.4 and will be evaluated. The efficiency score which differentiates between the control and experimental groups with the highest number of statistically significant different quantiles across all modules will be deemed the model with better sensitivity.

Information gain will be calculated as described in Chapter 3.4 and will be evaluated. The Information gain units for each efficiency score will be rated from 1 to 16 (1 - the most I.G. units to 16 - the least I.G. units) per module and the ratings will be aggregated across all the modules. This will be referred to as the Total Rating. Once all the rankings are aggregated for all 20 modules, the efficiency scores will be ranked again from 1 to 16 based on the total aggregated rankings (1 - the least aggregated rating achieved to 16 - the most aggregated rating achieved). This will be referred to as the Final Rank.The efficiency score which achieves Rank 1 will be deemed as the score which provides the highest information gain.

The classification models will be modelled as described in Chapter 3.4 and will be evaluated. For the models predicting the group, they will be evaluated on the Accuracy, Precision, Recall and F1 score of the models.

- **Accuracy** is a ratio of correctly predicted observations to the total observations.
- **Precision** is the ratio of correctly predicted positive observations to the total predicted positive observations.
- **Recall** is the ratio of correctly predicted positive observations to the all observations in actual class.
- **F1 Score** is the weighted average of Precision and Recall.

Accuracy is a great measure but only when the dataset is symmetric where values of false positive and false negatives are almost the same. In order to evaluate the performance of the model, the other parameters must be considered too. A high precision score gives more confidence to the model's capability to classify positive observations. Combining this with recall gives an idea of how many of the total positive observations the model is able to cover. A good model should have a good precision as well as a high recall.

For the models predicting the module, they will be evaluated on the Accuracy of the models. The model with the highest average accuracy score will be deemed the best performing model.

3.6 Strengths and Limitations of the Design

3.6.1 Strengths

The parabolic model has not been applied previously in any research. This is the first empirical application of the model and upon completion would be contributing to the field of education. This experiment design is based on the assumption that participants have no prior knowledge. Any student with prior knowledge of the topic being discussed in the module could end up achieving high performance with minimal effort and could result as a potential outlier, which will affect the analysis. A subjective measure of prior knowledge was requested from the participants to assess how much they knew before entering the class. PM model seeks to penalise learners with prior knowledge when calculating efficiency and is designed to spot outlier values, so that they can be accounted for when calculating the efficiency score.

The design framework with the models of efficiency and the measures employed to quantify the perceived mental effort and workload are easy to implement across any field and in turn, analyse. This empirical study follows the recommendations by Orru and Longo(2020) and undertakes statistical testing for small sample-size groups comparison by implementing techniques such as Shift Function.

Evaluating the two instructional conditions specified with the shared learning outcome for this design allow for their comparison as suggested by Jonassen(2009).

The use of two classification learning approaches makes the design framework robust. The design acknowledges both linear and non-linear data. The Logistic regression is a straight forward algorithm and they are not computationally intensive, while providing good interpret-ability. SVM are sophisticated and capable classifiers because they are able to carry out non-linear partitioning. The researcher does not have to transform the non-linear data themselves. They allow substantial flexibility for the decision boundaries, leading to better classification performance.

By including both variations of the multidimensional model formula by Tuovinen and Paas(2004) and Hoffman and Schraw (2010) in this experiment design, it can be investigated whether or not the inverse of mathematical operations will produce the same difference or quotient as argued by Hoffman and Schraw (2010).

3.6.2 Limitations

Although the intention was to divide the groups as evenly as possible for all classes, this was not always possible due to the participants present in the class on the day. It was also not possible to ensure that there were similar number of participants for all classes, as the size of the classes varied greatly. The experiment was carried out using a dataset with only 455 cases. With a limited dataset, it is difficult to achieve generalisation.

The use of subjective measures of mental effort by the participants themselves could lead to bias in the dataset and there is currently no way to address that in the experiment design. However, the advantage is that such a measure is easy to implement and analyse.

This experiment is limited to learners in third level education and instructors who deliver using direct instruction approach to learning. This design can be extended to accommodate different domains by using the same analysis.

The number of questions for MCQ in each module delivered lacked consistency. The varying number of questions across the different modules means that there could be inconsistent spread of the MCQ scores.

The use of SVM has its own disadvantages which can be a limitation to this research. They can be prone to over-fitting. The use of kernels to separate the non-linear data makes them difficult to interpret. SVMs are also very sensitive to the choice of the kernel parameters.

The outliers could be spotted by one model and not by another. There is a lack of proof to decide which models will and which model won't detect outliers in their calculations. The assumption is that all the models except for the parabolic model will not account for outliers in the efficiency score calculations.

Chapter 4

Results, Evaluation and Discussion

This chapter provides the reader with an outline of the results, its analysis and interpretation of the data from the implementation of the empirical study experiment as described in chapter 3. The results will be analysed and the models will be evaluated as described in **section 3.5**. The chapter discusses strengths and limitations of the results and evaluation from this experiment and problems encountered.

4.1 Results

Data preparation, exploration and analysis was conducted using R studio, primarily used in academics and research. R was chosen because it is an easier language to learn, statistical tests and models are readily available and can be easily used. The original dataset collected contained 25 variables and 455 observations. The details of the variables collected are provided in Table A.1. A good representation was observed from both control and experimental groups with 231 and 224 observations respectively across the 20 modules. Details of the module breakdown can be viewed in Table A.2.

The following variables were computed using data present in the original dataset, as described in **chapter 3**:

1. RAWNASA scores: Pre-MCQ and Post-MCQ
2. Standardised scores: MCQ, RSME and RAWNASA
3. Efficiency scores x 16

There were no "NA" values observed in the dataset for the important variables such as MCQ score, RSME scores and the six variables required to calculate the RAWNASA scores. Multiples instances of "NA" values were observed for "Knowledge" and "Motivation" variables (47 and 24 respectively). They were not required for calculating any of the efficiency scores or the RAWNASA scores. Therefore, these variables were removed from the dataset. No complete observations were removed from the dataset as a result of the "NA" values.

Uni-variate analysis was performed on the distributions of the important variables on the overall dataset and module subsets to determine normality and skewness of the variables. Standardised skewness was calculated for all variables to determine the normality of the variable and a majority of the variables returned as non-normal, with a standardised skewness score outside +/- 2 [1].The distributions were not observed to be bimodal or multimodal, where there are multiple peaks in the distribution.

Since the variables were non-normal and the sample size was very small, the assumptions for parametric tests were not met. Non-parametric tests were considered for inferential statistical tests. These include Spearman correlation test, Wilcoxon Signed Rank test and Kruskal-Wallis test. Any detected outliers in the dataset were not removed.

The overall increment of mental effort and workload has minimal or no effect on the performance of an individual as measured by the MCQ score. This is demonstrated

[1] (George & Mallery, 2010)

by the regression line along the scatter plots in Figure 4.1. This suggests that the two factors could be independent of each other and that the combination of these two factors will provide more insight by means of efficiency scores.

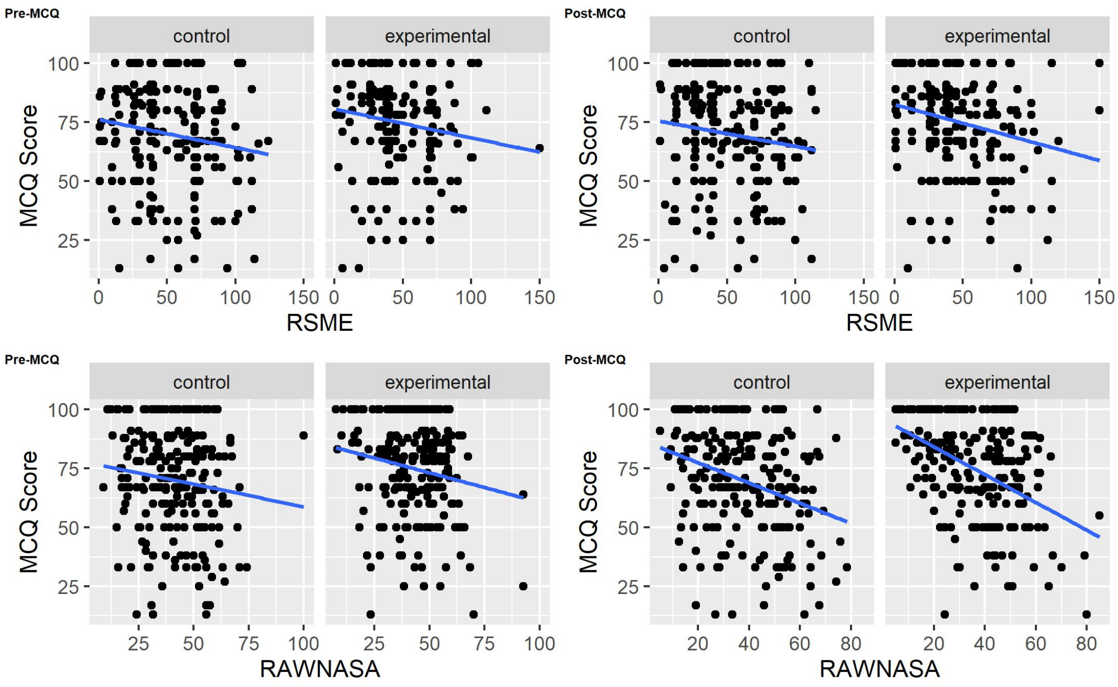

Figure 4.1: Overall relations between mental effort / workload (Pre-MCQ and Post-MCQ) and performance. The linear regression is represented by the blue line

Table 4.1 shows the mean, standard deviations (SD), median and the inter-quartile range (IQR) of the MCQ scores associated to each module and the related groups within each module. On average, the participants in the experimental group achieved higher MCQ scores compared to the control group. However, it can be observed that for some modules, the participants in the control group performed better and only by a small margin in some cases. Based on the median scores, the experimental group performed better in 10 of the modules and the control group performed better in three of the modules. It can also be observed that there were no differences in median values in the remaining seven modules, e.g. Modules C, F, J, K, Q, S & T. Outliers were detected in 10 modules as shown in Figure 4.2.

	control				**experimental**			
Module ID	**M**	**SD**	**Mdn**	**IQR**	**M**	**SD**	**Mdn**	**IQR**
A	71.50	22.12	77.50	15.75	75.33	14.37	73.00	22.50
B	82.35	17.38	89.00	16.50	89.00	13.91	100.00	22.00
C	45.40	20.89	50.00	25.00	53.86	17.31	50.00	12.50
D	65.65	22.04	67.00	27.75	84.11	12.06	83.50	11.00
E	76.21	24.18	88.00	13.00	54.82	25.27	50.00	31.50
F	37.80	14.44	38.00	18.75	32.22	11.04	38.00	13.00
G	77.00	16.85	75.00	19.00	64.50	14.17	64.00	25.00
H	79.38	13.71	78.00	13.75	86.25	7.78	89.00	11.00
I	58.30	22.69	58.50	29.75	74.92	21.89	83.00	24.50
J	85.33	19.22	100.00	20.00	88.00	16.56	100.00	20.00
K	77.00	8.22	71.00	15.00	76.63	7.76	71.00	15.00
L	69.29	15.33	71.00	14.50	77.78	10.65	86.00	15.00
M	66.00	22.71	71.00	21.25	87.33	11.30	86.00	14.00
N	50.17	27.73	58.50	41.75	62.00	21.01	67.00	17.00
O	52.63	17.85	50.00	25.50	58.29	18.25	58.50	17.00
P	71.43	15.74	60.00	20.00	75.56	16.67	80.00	20.00
Q	75.13	12.38	67.00	16.00	73.86	8.55	67.00	16.00
R	82.22	16.65	80.00	35.00	97.14	7.26	100.00	0.00
S	69.44	19.52	66.00	21.00	68.53	15.21	66.00	25.00
T	78.57	14.60	80.00	15.00	84.62	14.50	80.00	20.00

Table 4.1: Mean, SD, median and inter-quartile range of the MCQ scores grouped by control and experimental for each module

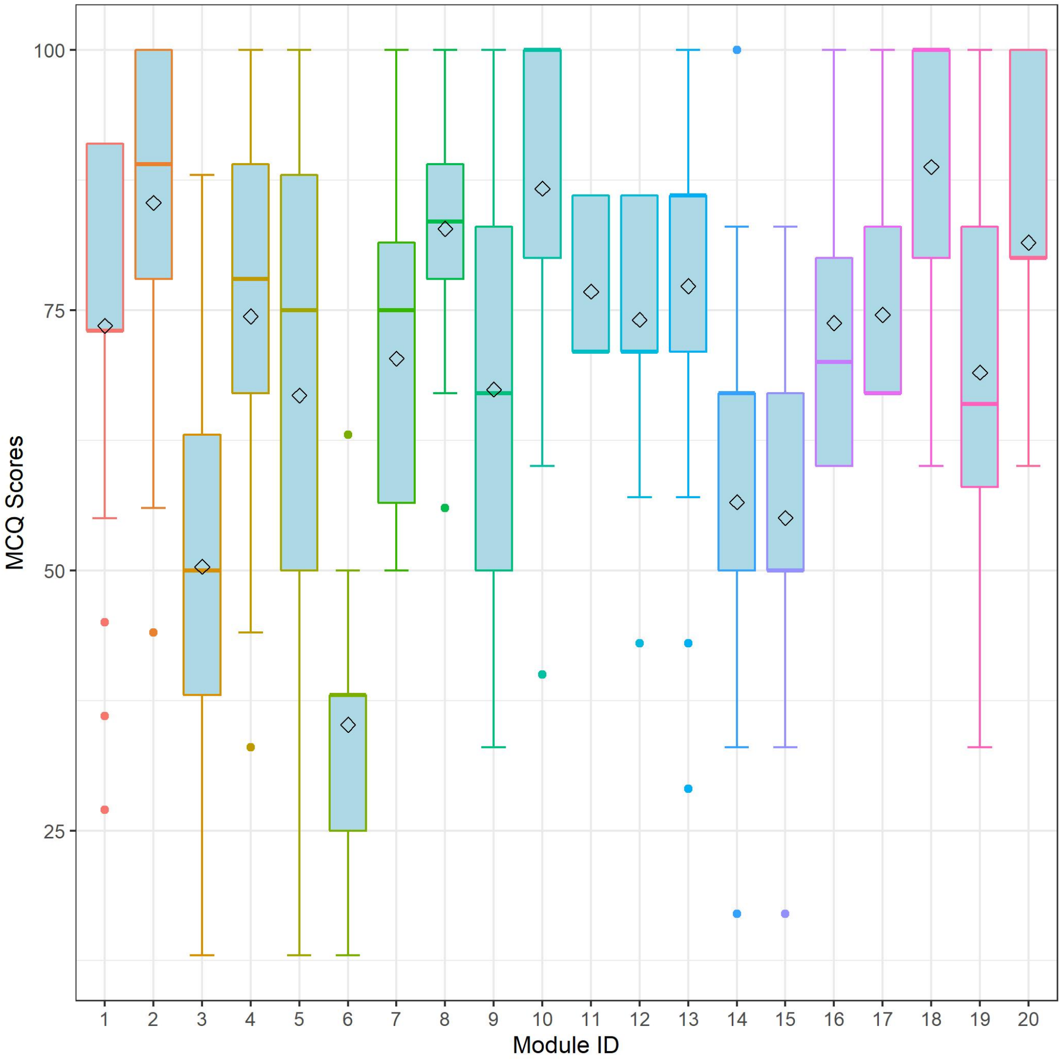

Figure 4.2: Boxplot of MCQ scores per module

Table 4.2 shows the mean, standard deviations (SD), median and the inter-quartile range (IQR) of the RSME scores (Pre-MCQ) associated to each module and the related groups within each module. The experimental group exerted less mental effort before the MCQ when compared to the control group in 10 modules. The control group exerted less mental effort before the MCQ when compared to the experimental group in nine modules. It can also be observed that there were no differences in median values between the control and experimental group in Module E. control group (Mdn = 40, IQR = 25.25 - 56.25) and experimental group (Mdn = 40, IQR = 36 - 70). Outliers were detected in nine modules as shown in Figure 4.3.

	control				**experimental**			
Module ID	**M**	**SD**	**Mdn**	**IQR**	**M**	**SD**	**Mdn**	**IQR**
A	49.36	29.14	38.00	36.50	56.53	35.31	50.00	41.50
B	55.75	27.20	61.50	38.75	37.38	21.56	38.00	15.25
C	39.80	35.95	27.00	21.00	47.86	22.51	40.00	20.00
D	45.90	32.39	40.00	38.25	37.89	25.89	38.00	39.75
E	39.36	22.76	40.00	31.00	48.73	25.91	40.00	34.00
F	54.10	30.08	47.50	24.00	38.56	23.59	38.00	13.00
G	50.00	26.07	50.00	34.50	33.38	9.98	38.00	10.00
H	41.25	26.26	40.00	34.25	56.75	22.32	49.00	30.50
I	42.60	34.52	39.00	62.25	35.08	22.19	37.50	14.75
J	49.00	25.99	38.00	24.00	51.73	31.44	40.00	27.50
K	35.20	31.77	12.00	58.00	35.13	23.26	39.00	36.25
L	35.71	25.22	30.00	11.00	45.56	16.64	40.00	9.00
M	55.38	17.36	58.50	29.00	35.56	24.77	26.00	20.00
N	61.17	32.07	55.50	35.75	56.86	11.94	60.00	18.00
O	51.32	30.23	38.00	45.00	44.14	21.39	39.00	32.00
P	69.86	24.67	72.00	16.00	63.22	19.97	68.00	20.00
Q	58.75	29.26	67.00	46.25	45.71	19.13	38.00	28.00
R	66.72	25.63	70.50	39.75	42.64	19.18	38.00	9.00
S	55.50	26.87	54.00	30.50	64.93	19.54	72.00	26.00
T	64.93	26.10	64.00	39.25	51.38	24.91	40.00	32.00

Table 4.2: Mean, SD, median and inter-quartile range of the Pre-MCQ RSME scores grouped by control and experimental for each module

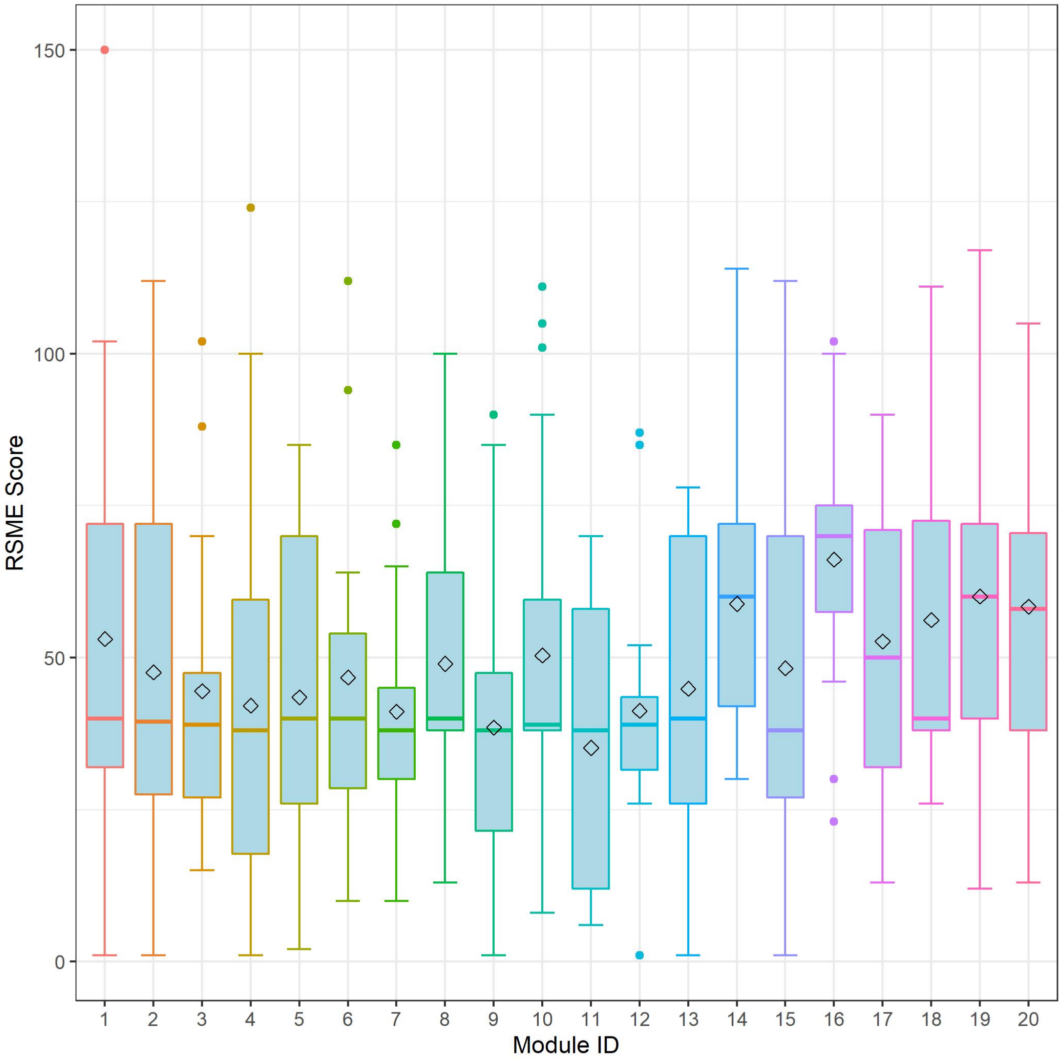

Figure 4.3: Boxplot of RSME scores (Pre-MCQ) per module

Table 4.3 shows the mean, standard deviations (SD), median and the inter-quartile range (IQR) of the RSME scores (Post-MCQ) associated to each module and the related groups within each module. The experimental group exerted less mental effort after the MCQ when compared to the control group in 13 modules. The control group exerted less mental effort before the MCQ when compared to the experimental group in six modules. It can also be observed that there were no differences in median values between the control and experimental group in Module G. control group (Mdn = 40, IQR = 28.5 - 70) and experimental group (Mdn = 40, IQR = 36 - 42.5). Outliers were detected in five modules as shown in Figure 4.4.

	control				**experimental**			
Module ID	**M**	**SD**	**Mdn**	**IQR**	**M**	**SD**	**Mdn**	**IQR**
A	51.21	26.38	49.00	36.50	50.13	26.61	45.00	41.00
B	60.60	28.91	71.00	42.75	41.69	25.99	38.00	30.25
C	53.80	37.70	32.00	45.00	79.57	24.69	80.00	34.00
D	31.50	20.66	30.00	26.00	38.83	34.76	27.00	23.25
E	50.93	27.40	43.50	38.25	47.18	22.55	40.00	36.00
F	56.30	34.39	60.00	44.00	82.56	26.81	85.00	28.00
G	47.00	22.91	40.00	41.50	42.38	15.02	40.00	6.50
H	47.63	21.23	45.00	26.25	69.50	36.71	69.00	18.50
I	48.40	35.19	40.00	59.50	36.83	15.05	39.00	18.00
J	43.87	34.01	27.00	34.00	60.73	35.05	60.00	43.50
K	47.40	24.97	38.00	33.00	58.13	30.30	40.00	50.00
L	38.14	17.72	45.00	25.00	24.00	17.06	25.00	17.00
M	50.75	20.19	41.50	18.75	42.33	13.38	40.00	7.00
N	73.83	27.04	71.50	26.25	43.86	21.87	40.00	26.00
O	50.79	25.98	40.00	39.50	43.57	29.18	39.00	23.25
P	65.14	30.17	74.00	50.50	52.89	26.68	60.00	32.00
Q	49.38	30.62	38.50	53.25	44.29	31.96	38.00	40.50
R	62.06	26.76	65.00	43.50	31.93	22.38	26.00	27.75
S	64.38	25.52	60.00	43.25	57.40	19.62	70.00	29.50
T	52.29	27.84	49.00	45.50	35.54	20.23	38.00	14.00

Table 4.3: Mean, SD, median and inter-quartile range of the Post-MCQ RSME scores grouped by control and experimental for each module

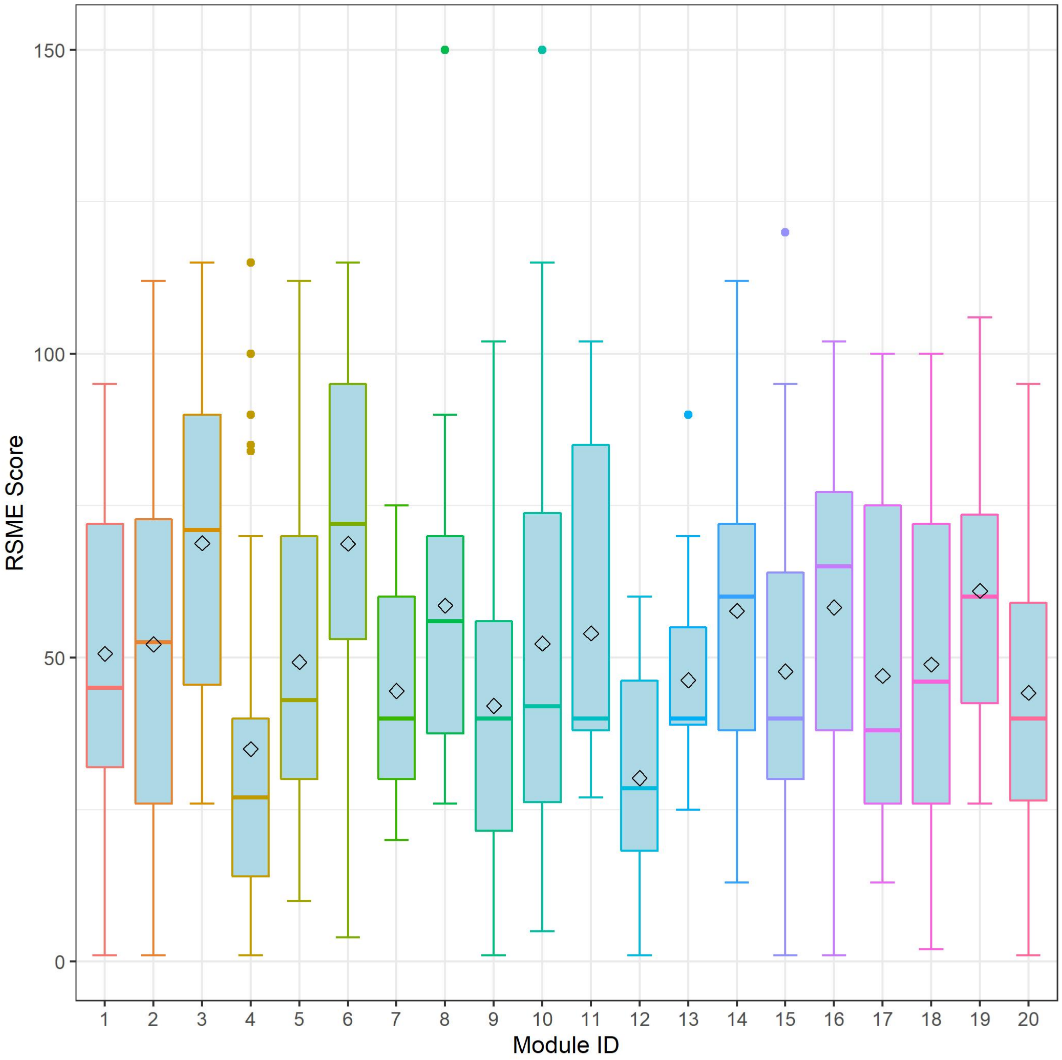

Figure 4.4: Boxplot of RSME scores (Post-MCQ) per module

Table 4.4 shows the mean, standard deviations (SD), median and the inter-quartile range (IQR) of the RAWNASA scores (Pre-MCQ)associated to each module and the related groups within each module. The experimental group exerted less mental workload before the MCQ when compared to the control group in eight modules. The control group exerted less mental workload after the MCQ when compared to the experimental group in 11 modules. It can also be observed that there were no differences in median values between the control and experimental group in Module P. control group (Mdn = 45, IQR = 39.58 - 63.75) and experimental group (Mdn = 45, IQR = 35 - 51.67). Outliers were detected in seven modules as shown in Figure 4.5.

	control				**experimental**			
Module ID	**M**	**SD**	**Mdn**	**IQR**	**M**	**SD**	**Mdn**	**IQR**
A	47.14	12.03	48.75	12.08	47.33	17.40	48.33	20.42
B	45.79	17.61	48.75	21.25	45.16	10.93	48.33	9.17
C	35.83	13.16	39.17	22.50	43.69	11.73	41.67	11.25
D	37.50	12.56	32.92	10.42	42.36	13.77	46.67	16.88
E	39.94	14.03	41.25	6.67	46.21	10.64	49.17	15.00
F	42.92	8.37	44.17	13.13	47.96	21.23	38.33	12.50
G	37.14	16.73	35.00	21.25	38.65	12.36	38.75	13.75
H	34.79	15.68	32.50	20.63	52.71	7.63	54.17	8.96
I	41.25	19.55	44.17	27.08	30.35	8.35	29.17	6.67
J	35.39	15.89	39.17	20.00	39.61	17.84	40.83	20.42
K	33.17	5.51	34.17	9.17	38.54	6.12	37.08	7.71
L	35.48	13.07	35.00	18.75	49.72	12.08	53.33	5.00
M	46.04	18.99	51.67	24.79	41.76	16.87	47.50	15.00
N	55.83	8.74	55.42	4.58	53.21	7.88	55.00	9.17
O	34.34	14.32	30.83	19.58	39.70	16.27	37.50	20.42
P	50.12	14.60	45.00	24.17	43.89	9.84	45.00	16.67
Q	42.29	8.88	42.50	11.67	46.55	9.72	51.67	15.00
R	47.87	11.55	50.00	19.79	35.00	12.48	36.67	11.25
S	47.29	10.79	51.67	14.58	45.50	11.00	45.83	8.75
T	37.92	11.39	35.00	15.42	42.31	13.90	47.50	20.00

Table 4.4: Mean, SD, median and inter-quartile range of the Pre-MCQ RAWNASA scores grouped by control and experimental for each module

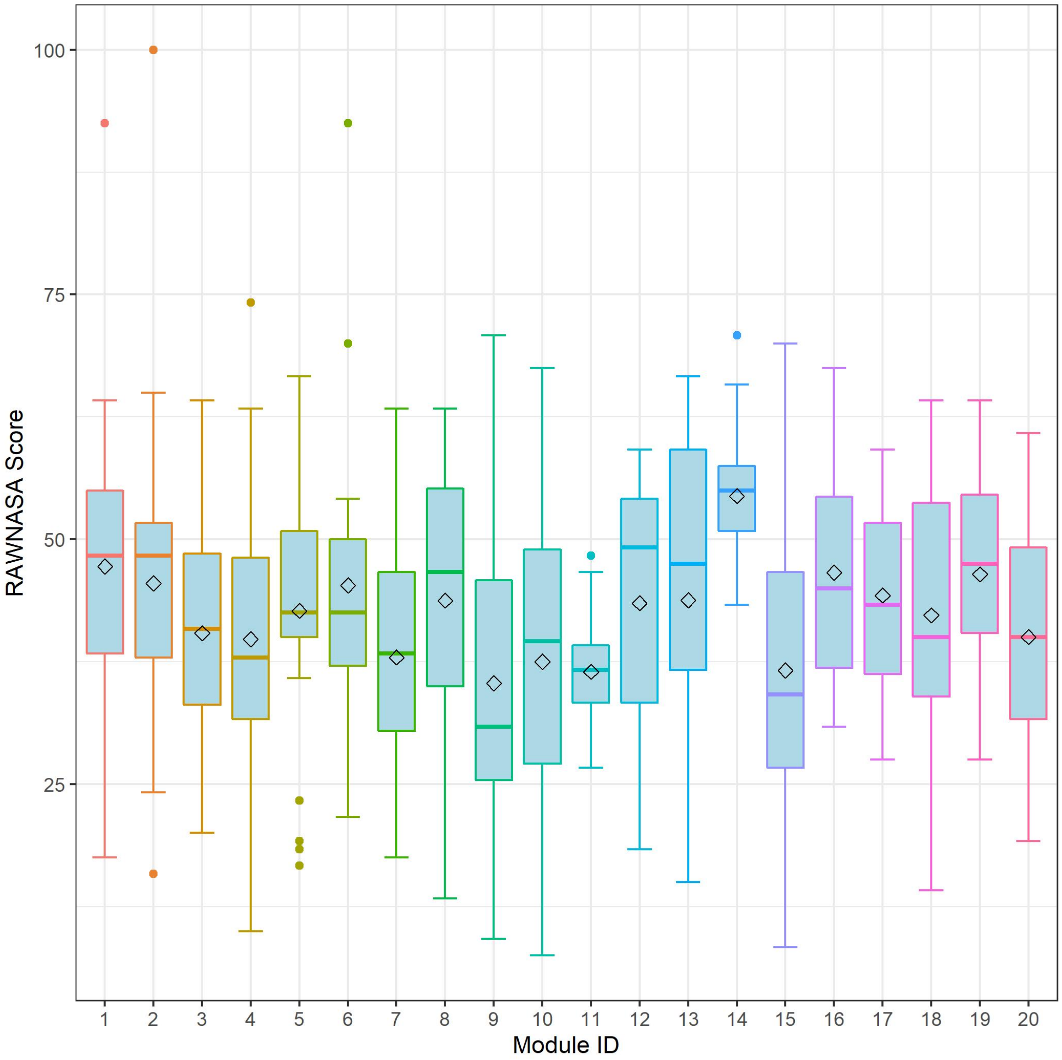

Figure 4.5: Boxplot of RAWNASA scores (Pre-MCQ) per module

Table 4.5 shows the mean, standard deviations (SD), median and the inter-quartile range (IQR) of the RAWNASA scores (Post-MCQ) associated to each module and the related groups within each module. The experimental group exerted less mental workload after the MCQ when compared to the control group in 13 modules. The control group exerted less mental workload after the MCQ when compared to the experimental group in six modules. It can also be observed that there were no differences in median values between the control and experimental group in Module T. control group (Mdn = 25, IQR = 20.21 - 38.54) and experimental group (Mdn = 25, IQR = 20 - 43.33). Outliers were detected in two modules as shown in Figure 4.6.

	control				**experimental**			
Module ID	**M**	**SD**	**Mdn**	**IQR**	**M**	**SD**	**Mdn**	**IQR**
A	41.37	20.68	46.25	28.33	41.00	19.08	39.17	22.50
B	40.25	15.40	39.17	10.42	37.66	15.92	45.00	28.54
C	42.00	11.17	45.83	16.67	47.86	5.95	50.00	8.75
D	32.29	16.52	32.08	17.29	29.58	15.26	27.08	24.58
E	40.00	17.66	41.67	27.92	42.65	12.39	44.17	14.17
F	46.08	14.00	45.83	21.25	58.15	15.60	55.83	20.00
G	32.86	17.79	26.67	31.67	33.13	13.44	32.08	16.67
H	30.73	14.70	29.17	9.58	49.17	8.52	49.17	7.29
I	33.33	14.71	34.17	12.71	29.17	12.50	29.17	12.71
J	30.11	19.78	23.33	16.25	29.72	16.09	28.33	22.50
K	30.00	8.27	29.17	8.33	34.58	13.67	34.17	19.37
L	34.17	8.31	32.50	8.33	16.94	5.51	16.67	4.17
M	45.52	19.23	50.42	31.67	33.43	14.03	28.33	22.50
N	40.42	14.08	43.33	16.88	39.64	8.51	40.00	10.00
O	38.68	18.30	35.83	30.00	41.31	17.89	42.92	23.75
P	39.05	17.49	38.33	25.42	37.59	16.44	37.50	22.50
Q	30.94	12.19	30.42	12.71	42.14	12.11	41.67	13.75
R	41.44	11.61	45.42	16.67	18.33	11.96	14.58	15.83
S	45.26	14.72	52.50	15.42	36.89	10.53	36.67	15.42
T	29.35	13.48	25.00	18.33	31.60	15.02	25.00	23.33

Table 4.5: Mean, SD, median and inter-quartile range of the Post-MCQ RAWNASA scores grouped by control and experimental for each module

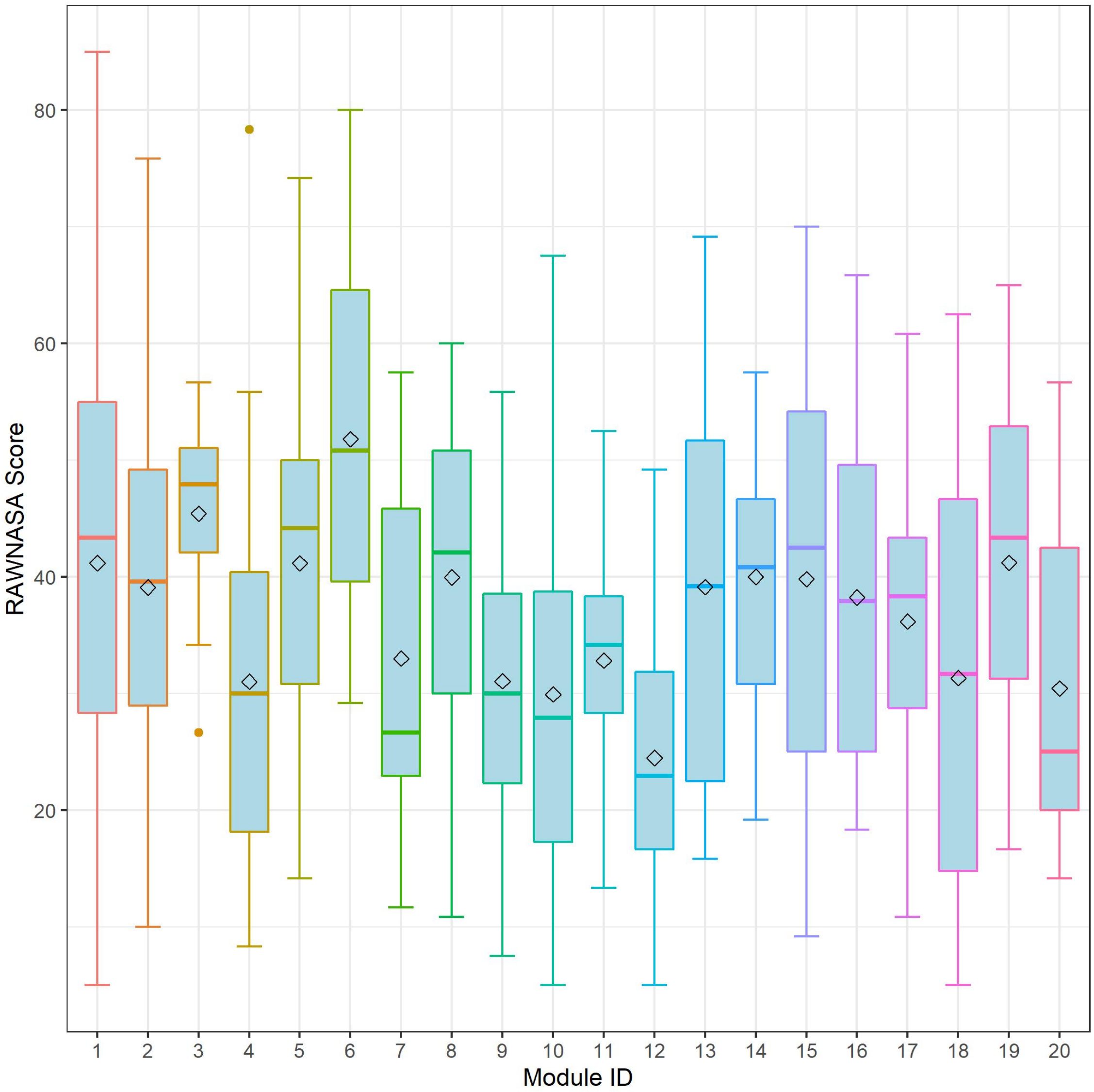

Figure 4.6: Boxplot of RAWNASA scores (Post-MCQ) per module

A Wilcoxon signed rank test was performed to compare the medians of control and experimental groups. Table 4.6 shows the resulting p-values of the Wilcoxon signed rank tests computed for the following variables: MCQ score, RSME scores and RAWNASA scores.

Despite there being a difference in performance between the control and experiment groups on 13 of the modules as shown in Table 4.1, the results of the Wilcoxon signed rank test are statistically significant for only five of the modules as shown in

Table 4.6. All other p-values were above the α level. Given the dynamics of the third level education modules, and the variations in the participants' approach to the modules, this is expected.

Despite there being a difference in RSME Score (Pre-MCQ) between the control and experiment groups on 19 of the modules as shown in Table 4.2, the results of the Wilcoxon signed rank test are statistically significant for only one module as shown in Table 4.6 - Module R (W=193.5, p = .01, r = 0.46). All other p-values were above the α level.

Despite there being a difference in RSME Score (Post-MCQ) between the control and experiment groups on 19 of the modules as shown in Table 4.3, the results of the Wilcoxon signed rank test are statistically significant for only one module as shown in Table 4.6 - Module R (W=204.5, $p < .001$, r = 0.53). All other p-values were above the α level.

Despite there being a difference in RAWNASA Score (Pre-MCQ) between the control and experiment groups on 19 of the modules as shown in Table 4.4, the results of the Wilcoxon signed rank test are statistically significant for only three of the modules as shown in Table 4.6 - Module H (W=10, p = .02, r = 0.58), Module L (W=11.5, p = .04, r = 0.53) and Module R (W=188.5, p = .02, r = 0.42). All other p-values were above the α level.

Despite there being a difference in RAWNASA Score (Post-MCQ) between the control and experiment groups on 19 of the modules as shown in Table 4.5, the results of the Wilcoxon signed rank test are statistically significant for only three of the modules as shown in Table 4.6 - Module H (W=10, p = .02, r = 0.58), Module L (W=63, $p < .001$, r = 0.84) and Module R (W=229.5, $p < .001$, r = 0.70). All other p-values were above the α level.

		RSME		RAWNASA	
Module ID	MCQ Scores	Pre-MCQ	Post-MCQ	Pre-MCQ	Post-MCQ
A	**.02**	.60	1	.83	.78
B	.19	.07	.05	.87	.97
C	.80	.22	.20	.37	.57
D	**.01**	.48	.90	.05	.60
E	**.04**	.35	.96	.21	.83
F	.39	.14	.09	.84	.11
G	.17	.27	1	.87	1
H	.29	.31	.16	**.02**	**.02**
I	.10	.74	.57	.15	.41
J	.75	.60	.11	.55	.85
K	1	.94	.34	.27	.56
L	.26	.09	.18	**.04**	**< .001**
M	**.04**	.06	.38	.53	.16
N	.60	1	.10	.62	.53
O	.42	.70	.48	.32	.81
P	.64	.34	.39	.43	.87
Q	1	.35	.68	.52	.20
R	**.01**	**.01**	**< .001**	**.02**	**< .001**
S	.93	.10	.45	.55	.07
T	.29	.15	.14	.38	.75

Table 4.6: p-values at $\alpha = .05$ of Wilcoxon Rank Sum test of the MCQ score, Pre-MCQ RSME score, Post-MCQ RSME score, Pre-MCQ RAWNASA score, Post-MCQ RAWNASA score

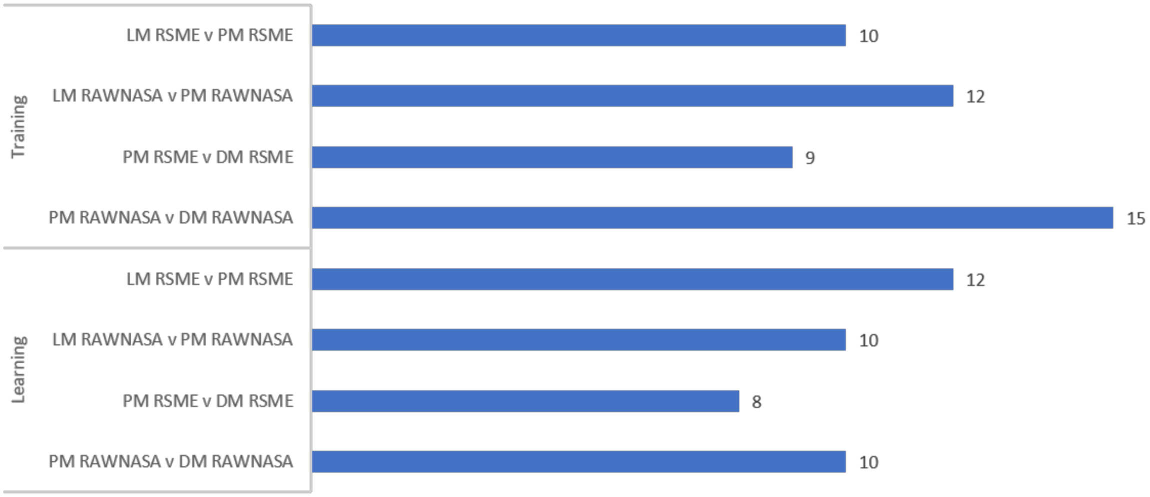

Figure 4.7: Concurrent validity: Number of statistically significant efficiency score correlation pairs

Efficiency scores (training and learning efficiency) of all the two-dimensional models (LM, PM and DM) were paired up with each other and the concurrent validity for the models of efficiency were calculated using Spearman correlation test across all 20 modules. Both efficiency scores utilising RSME and RAWNASA were examined for the concurrent validity of the models. Figure 4.7 illustrates the number of statistically significant correlations between the efficiency score pairs. Among the training efficiency pairs, the efficiency score pair between PM and DM using RAWNASA mental workload score resulted in the highest number of significant correlations (15) across 20 modules. The efficiency score pair between PM and DM using RSME mental effort score resulted in the lowest number of significant correlations (9) across the 20 modules. Among the learning efficiency pairs, the efficiency score pair between LM and PM using RSME mental effort score resulted in the highest number of significant correlations (12) across the 20 modules. The efficiency score pair between PM and DM using RSME mental effort score resulted in the lowest number of significant correlations (8) across the 20 modules. Out of 80 possible training efficiency score pairs, 40 were statistically significant. Out of 80 possible learning efficiency score pairs, 46 were statistically significant.

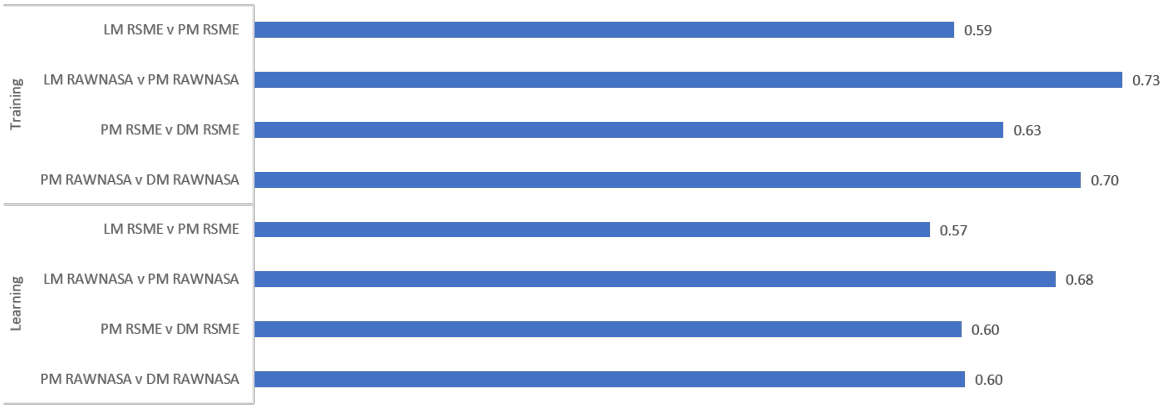

Figure 4.8: Concurrent validity: Average r_s for all statistically significant efficiency score correlation pairs

Figure 4.8 illustrates the average spearman's rho (r_s) values between the efficiency score pairs for all the statistically significant pairs across the 20 modules. Among the training efficiency pairs, the efficiency score pair between LM and PM using RAWNASA mental workload score has the highest average correlation (r_s = .73). The efficiency score pair between LM and PM using RSME mental effort score has the lowest average correlation (r_s = .59). Among the learning efficiency pairs, the efficiency score pair between LM and PM using RAWNASA mental workload score has the highest average correlation (r_s = .68). The efficiency score pair between LM and PM using RSME mental effort score has the lowest average correlation (r_s = .57). It is interesting to note that all the efficiency score pairs have an average correlation of r_s > .50.

The correlation between the training and learning efficiency scores of each two-dimensional model (LM, PM and DM) were examined using Spearman's correlation test to determine the discriminant validity of the models of efficiency across all 20 modules. Both efficiency scores utilising RSME and RAWNASA were examined for the discriminant validity of the models. Figure 4.9 illustrates the number of statistically significant correlations between the efficiency score pairs. Among the likelihood model (LM) of efficiency, the correlation pair of the efficiency scores utilising RAW-

NASA mental workload score resulted in the highest number of statistically significant correlations (16) across 20 modules. The correlation pair of the efficiency scores utilising RSME mental effort score resulted in the lowest number of statistically significant correlations (14) across 20 modules. Among the parabolic model (PM) of efficiency, both the correlation pair of the efficiency scores utilising RAWNASA mental workload score and RSME mental effort resulted in the same number of statistically significant correlations (13) across 20 modules. Among the deviational model (DM) of efficiency, the correlation pair of the efficiency scores utilising RSME mental effort score resulted in the highest number of statistically significant correlations (19) across 20 modules. The correlation pair of the efficiency scores utilising RAWNASA mental workload score resulted in the lowest number of statistically significant correlations (16) across 20 modules.

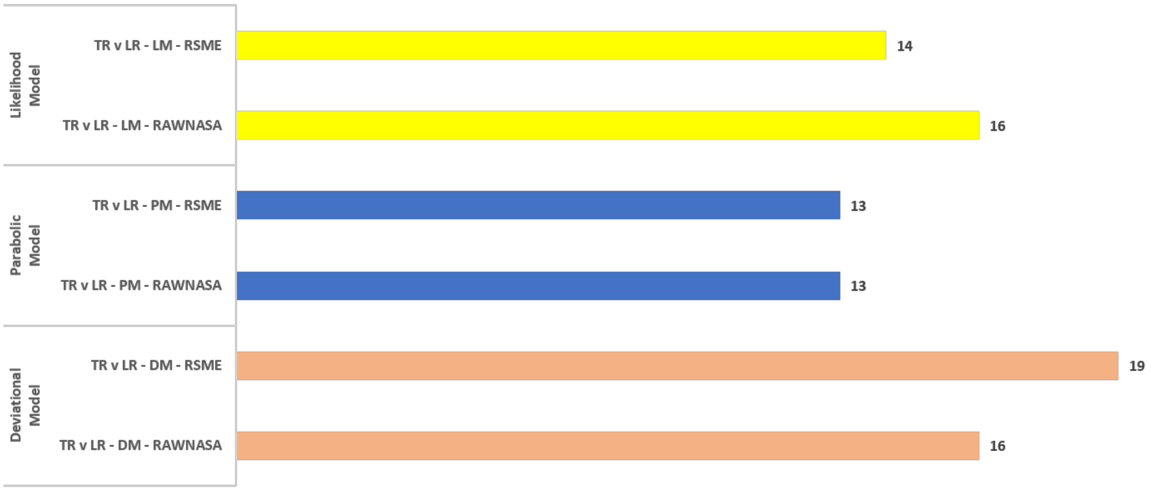

Figure 4.9: Discriminant validity: Number of statistically significant correlation pairs (training efficiency - learning efficiency) per model of efficiency

Figure 4.10 illustrates the average spearman's rho (r_s) values between the training and learning efficiency score correlations for all the statistically significant pairs across the 20 modules. It can be observed that the pairs of efficiency scores utilising RAWNASA mental workload score have a higher average correlation when compared to the pairs of efficiency scores utilising RSME mental effort. The DM efficiency score pair utilising RAWNASA has the highest average correlation ($r_s = .79$), while the PM efficiency score pair utilising RSME has the lowest average correlation ($r_s = .62$). All the efficiency score pairs have an average correlation of $r_s > .60$.

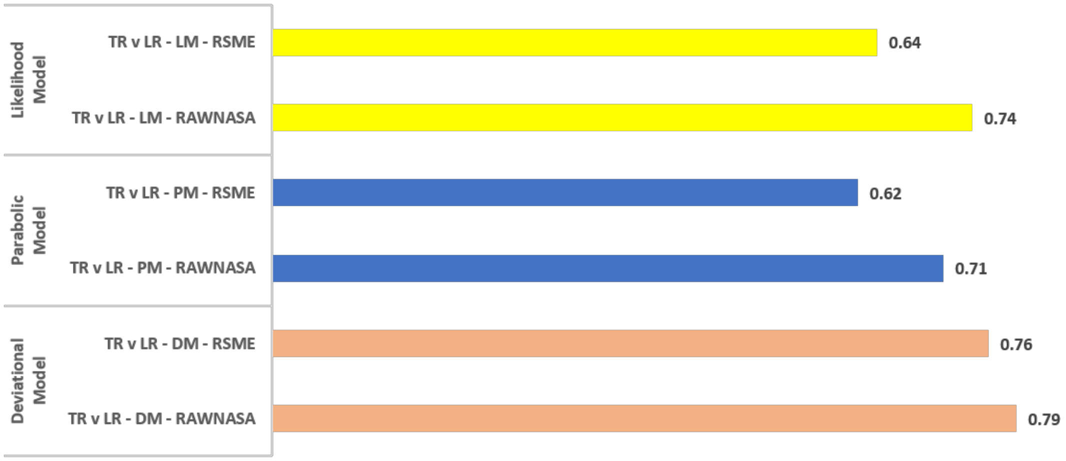

Figure 4.10: Discriminant validity: Average r_s for all statistically significant correlation pairs (training efficiency - learning efficiency) per model of efficiency

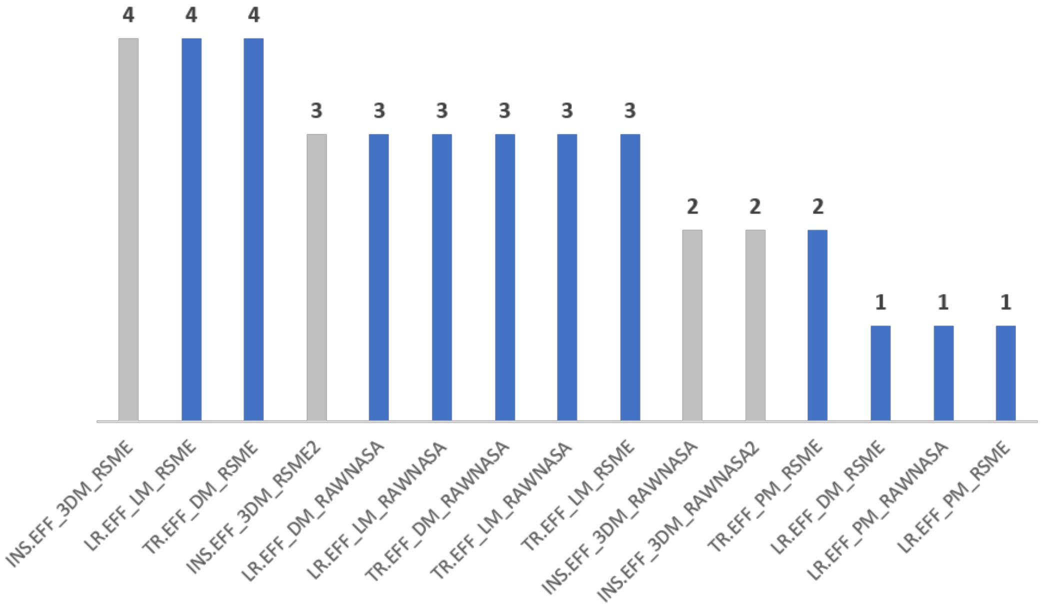

Figure 4.11: Known Groups validity: Number of instances with statistically significant differences between the control and experimental groups per efficiency score

Kruskal Wallis test along with the Post-Hoc Dwass-Steele-Critchlow-Fligner all-pairs test were conducted to test for difference between the control and experimental groups for a given efficiency score. The test was conducted on all 16 efficiency scores across all 20 modules.

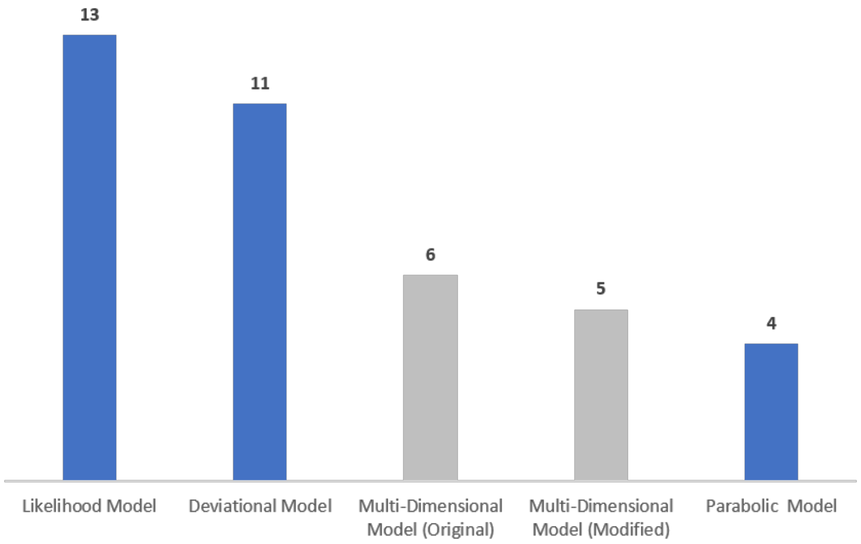

Figure 4.12: Known Groups validity: Number of instances with statistically significant differences between the control and experimental groups per model of efficiency

Figure 4.11 illustrates the number of statistically significant differences between the control and experimental groups grouped by the efficiency score. Out of 320 tests, it can be observed that only 39 produced statistically significant differences. The scores which provide the most number of statistical differences utilise RSME mental effort scores, with 4 differences each. Efficiency scores based on the likelihood model provided the highest number of statistical differences between the groups with 13, followed by efficiency scores based on the deviational model with 11 differences. Efficiency scores based on the parabolic model had the lowest result with 4 significant differences. The Shift function was employed to assess the sensitivity of the efficiency scores to differentiate between the control and the experimental groups. The shift function was employed on all 16 efficiency scores across all 20 modules. The shift function is a good visual tool to compare the differences between the two groups and determine how and by how much the two distributions differ. The shift function shows quantile differences between the control and experimental group, as a function of the control group. The quantiles were accepted as different if the confidence interval line shown on the plot was above or below ”0” on the X-axis. The confidence interval line touching the 0 line was not accepted as statistically different. Figure 4.13 shows the number of statistically significant different quantiles resulting from the shift function grouped by the efficiency scores. It can be observed that the efficiency score which differentiated between the control and experimental group the most from 20 modules was TR.EFF_LM_RSME with 11 quantiles, out of 60 possible quantiles (3 quantiles per module). LR.EFF_LM_RSME efficiency score is quite consistent in differentiating between the two groups based on the results from the Kruskal-Wallis tests and the shift function. TR.EFF_PM_RAWNASA is not represented in Figure 4.13 because it did not show any statistically significant quantile differences.

Efficiency scores based on the parabolic model was the worst performing performing two-dimensional model of efficiency with a total of 14 quantile differences out of 80 quantiles (20 module x 4 scores) which represents 17.5%. Efficiency scores based on likelihood model performed the best amongst the two-dimensional models with 37 quantiles differences out of 80 (46.25%), followed by the deviational model efficiency scores with 24 quantiles differences out of 80 (30%). Efficiency scores based on multidimensional model (Modified) performed the best amongst the three-dimensional models with 12 quantiles differences out of 40 quantiles (20 module x 2 scores) (30%), followed by multidimensional model (original) efficiency scores with 10 quantile differences out of 40 quantiles (25%). Considering the percentages of statistically significant different quantiles, the parabolic model was the least effective at differentiating between the two groups.

The statistically significant quantile differences were explored further to determine whether the shift was in favour of the control group or the experimental group. Figure 4.14 shows the breakdown grouped by the efficiency scores. The bars in orange represents the shift in favour of the control group, meaning that the control group had the better efficiency and the bars in purple represents the shift in favour of the experimental group, meaning that the experimental group had the better efficiency scores in the distribution.

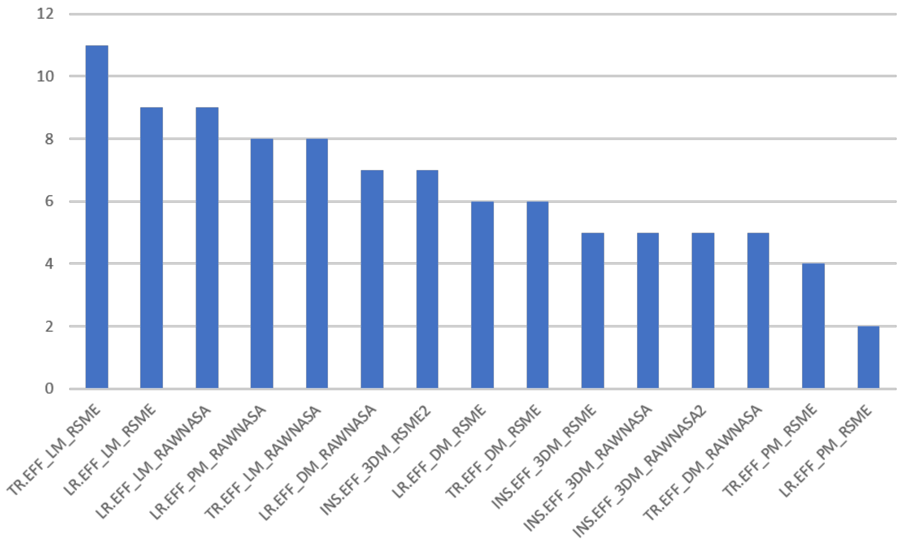

Figure 4.13: Shift Function: Number of statistically significant different quantiles across the 20 modules per efficiency score

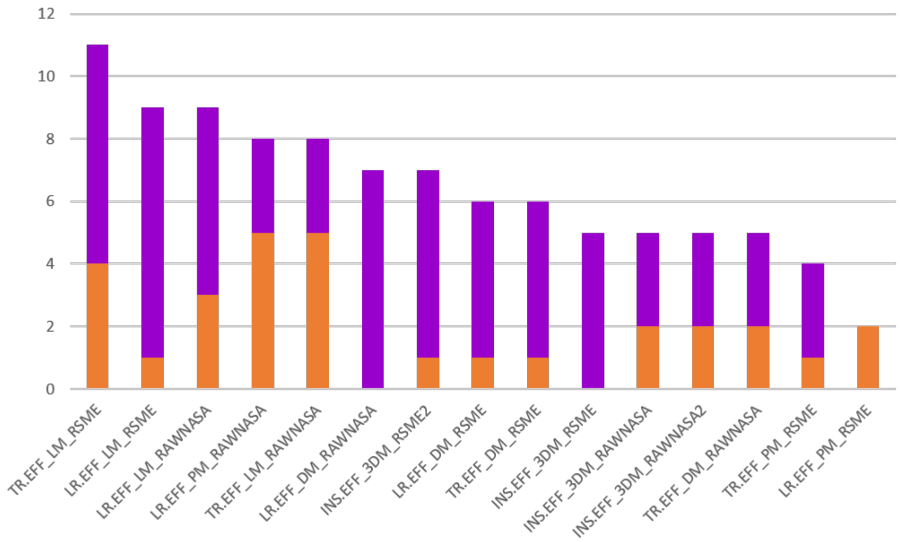

Figure 4.14: Shift Function: Number of quantile differences in favour of control group (orange) v experimental group (purple)

It can be observed from Figure 4.14 that in a majority of the cases, the quantile shifts were in favour of the experimental group with a ratio of 67:30. This would indicate that the experimental group in general had the better efficiency scores across the 20 modules.

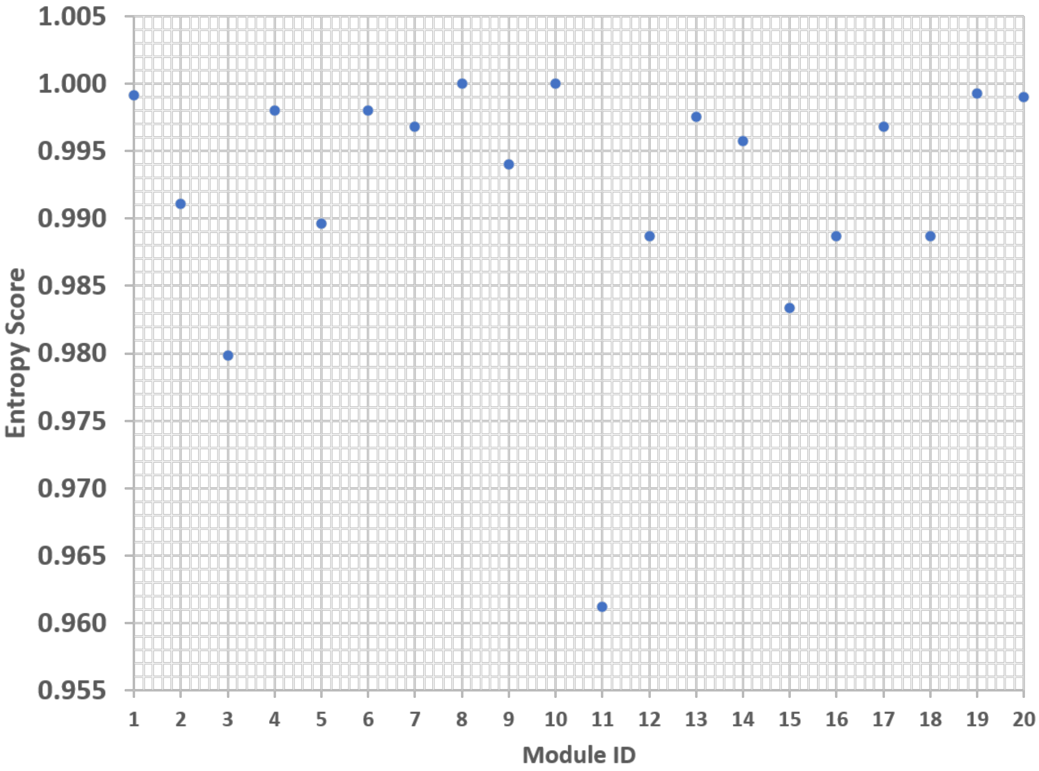

Figure 4.15: Entropy scores calculated per module for the variable ”group”

Entropy scores were calculated for the variable ”group” for each module’s data. Entropy scores show how pure the ”group” feature is. It provides an indication about the amount of knowledge that can be obtained about the group variable. Knowledge in this context refers to the certainty of drawing a specific observation at random from the dataset.

The higher the knowledge, the lower the entropy score. Figure 4.15 shows the calculated entropy scores. It can be observed that the module which has the lowest entropy is Module 11. There are two modules with the highest entropy with a value of 1: Module 8 and Module 10.

Efficiency Score	**Total Rating**	**Average I.G. Units**	**Rank**
LR.EFF_DM_RAWNASA	126	0.254	1
INS.EFF_3DM_RSME2	144	0.222	2
INS.EFF_3DM_RSME	149	0.225	3
LR.EFF_DM_RSME	149	0.220	4
TR.EFF_DM_RSME	149	0.218	5
TR.EFF_PM_RSME	152	0.218	6
INS.EFF_3DM_RAWNASA2	163	0.169	7
TR.EFF_LM_RAWNASA	167	0.195	8
TR.EFF_PM_RAWNASA	170	0.176	9
TR.EFF_DM_RAWNASA	172	0.222	10
LR.EFF_PM_RSME	176	0.187	11
LR.EFF_PM_RAWNASA	185	0.190	12
INS.EFF_3DM_RAWNASA	190	0.186	13
LR.EFF_LM_RAWNASA	190	0.171	14
LR.EFF_LM_RSME	193	0.167	15
TR.EFF_LM_RSME	216	0.126	16

Table 4.7: Ranking of efficiency scores based on the Information gain (I.G.) units calculated

Information gain was calculated for each of the efficiency scores to explore which efficiency score provides the most information about whether or not an observation belongs to the control or experimental group. By obtaining the information gain, it is possible to determine which one of the efficiency scores provides the ”purest” segmentation with respect to the groups. Table 4.7 provides a summary of all the information gain units (I.G. units) calculated for all the efficiency scores across the 20 modules and rated and ranked as specified in section 3.5. It can be observed that LR.EFF_DM_RAWNASA achieved the lowest aggregated rating and the best rank among the 16 efficiency scores, with a total rating of 126 and an average I.G. unit of 0.254. This means that this efficiency score provides the greatest information gain.

The efficiency score that provides the least information gain is TR.EFF_LM_RSME with a total aggregated rating of 216 and average I.G. unit of 0.126.

Classifier models were built, trained and tested using training and test data sets with 80:20 split ratio, partitioned from the overall dataset. Proportionate stratified sampling was used to ensure that both the training and test data sets had representative samples from all 20 modules. The training data set was used to train the classifier models and then the models were tested using the test data set.

Predictor	Accuracy	Precision	Recall	F-1 Score
INS.EFF_3DM_RSME	0.62	0.57	0.77	0.66
LR.EFF_LM_RSME	0.59	0.78	0.20	0.32
TR.EFF_DM_RAWNASA	0.58	0.55	0.69	0.61
LR.EFF_DM_RSME	0.58	0.54	0.71	0.62
INS.EFF_3DM_RSME2	0.58	0.83	0.14	0.24
LR.EFF_LM_RAWNASA	0.57	0.55	0.46	0.50
TR.EFF_LM_RSME	0.57	0.80	0.11	0.20
TR.EFF_PM_RAWNASA	0.57	0.53	0.80	0.64
INS.EFF_3DM_RAWNASA	0.57	0.53	0.77	0.63
LR.EFF_DM_RAWNASA	0.55	0.54	0.43	0.48
TR.EFF_DM_RSME	0.55	0.52	0.71	0.60
TR.EFF_LM_RAWNASA	0.54	0.56	0.14	0.23
LR.EFF_PM_RSME	0.54	0.67	0.06	0.11
LR.EFF_PM_RAWNASA	0.53	0.50	0.03	0.05
TR.EFF_PM_RSME	0.53	0.50	0.74	0.60
INS.EFF_3DM_RAWNASA2	0.53	0.50	0.29	0.36

Table 4.8: Accuracy, Precision, Recall and F-1 Scores for the binomial logistic regression classifiers grouped by efficiency score predictor

Table 4.8 provides a breakdown of the evaluation metrics for binomial logistic regression models built with the efficiency scores as the sole predictor variable to predict the groups to which an observation belongs. It can be observed that the classifier model built using the INS.EFF_3DM_RSME efficiency score provides the best accuracy (62%) and F-1 score (66%), which is a weighted average of precision and recall metrics. The classifier built using LR.EFF_LM_RSME was second with 59% accuracy and a F-1 score of 32%. Classifier built using TR.EFF_DM_RAWNASA was third with 58% accuracy, but had a much higher F-1 score of 61%. Taking all the metrics into consideration, TR.EFF_DM_RAWNASA classifier is considered the second best model. Classifiers built with efficiency scores based on the parabolic model had consistent accuracy scores ranging between 53% and 57%, however had varied F-1 scores ranging between 5% and 64%. It is also important to note that classier built using LR.EFF_PM_RAWNASA had the lowest F-1 score (5%) of all the binomial logistic regression classifiers. Classifier built using the INS.EFF_3DM_RAWNASA2 efficiency score performed the worst 53% accuracy, however had a F-1 score of 36% which can be considered in the middle.

Table 4.9 provides a breakdown of the evaluation metrics for support vector machine classifiers using radial kernel built with the efficiency scores as the sole predictor variable to predict to predict the groups to which an observation belongs. The Radial kernel results were selected for discussion as it provided the best overall average of accuracy. Evaluation metrics for the other types of kernels can be found in the Table B.1, Table B.2 and Table B.3. It can be observed from Table 4.9 that the classifier model built using the INS.EFF_3DM_RSME efficiency score provides the best accuracy (61%) and a F-1 score of 64%.

The accuracy and F-1 score produced by INS.EFF_3DM_RSME classifier model is consistent between the binomial logistic regression (Table 4.8) and support vector machine (radial kernel) algorithms. The classifier built using INS.EFF_3DM_RAWNASA2 was second with 59% accuracy and a F-1 score of 53%. Classifier built using LR.EFF_DM_RSME was third with 58% accuracy, but had a higher F-1 score of 64%. Classifiers built with efficiency scores based on the parabolic model had consistent accuracy scores of 55% on average, however had varied F-1 scores ranging between 33% and 55%. Classifier built using the INS.EFF_3DM_RSME2 efficiency score performed the worst 39% accuracy and had a F-1 score of 35% which was the third lowest.

Predictor	**Accuracy**	**Precision**	**Recall**	**F-1 Score**
INS.EFF_3DM_RSME	0.61	0.71	0.57	0.63
INS.EFF_3DM_RAWNASA2	0.59	0.49	0.59	0.53
LR.EFF_DM_RSME	0.58	0.77	0.54	0.64
LR.EFF_PM_RAWNASA	0.55	0.54	0.48	0.52
LR.EFF_LM_RAWNASA	0.55	0.34	0.55	0.42
TR.EFF_DM_RAWNASA	0.55	0.69	0.52	0.59
TR.EFF_PM_RAWNASA	0.55	0.69	0.46	0.55
TR.EFF_PM_RSME	0.55	0.43	0.54	0.48
TR.EFF_LM_RAWNASA	0.55	0.57	0.53	0.55
LR.EFF_DM_RAWNASA	0.54	0.74	0.51	0.60
INS.EFF_3DM_RAWNASA	0.53	0.26	0.50	0.34
LR.EFF_PM_RSME	0.53	0.29	0.38	0.33
TR.EFF_DM_RSME	0.50	0.54	0.51	0.53
TR.EFF_LM_RSME	0.50	0.83	0.48	0.61
LR.EFF_LM_RSME	0.42	0.74	0.43	0.55
INS.EFF_3DM_RSME2	0.39	0.34	0.35	0.35

Table 4.9: Accuracy, Precision, Recall and F-1 Scores for the Support vector machine classifiers using radial kernel grouped by efficiency score predictor

Table 4.10 provides a breakdown of the accuracy scores for multinomial logistic regressions models built with the efficiency scores as the sole predictor variable to predict to predict the modules to which an observation belongs. It can be observed that the classifiers built using the LR.EFF_PM_RAWNASA and TR.EFF_PM_RAWNASA effi-

Predictor	**Accuracy**
LR.EFF_PM_RAWNASA	0.19
TR.EFF_PM_RAWNASA	0.15
LR.EFF_LM_RSME	0.12
TR.EFF_LM_RAWNASA	0.11
LR.EFF_LM_RAWNASA	0.11
TR.EFF_LM_RSME	0.09
INS.EFF_3DM_RSME2	0.09
TR.EFF_PM_RSME	0.08
LR.EFF_DM_RAWNASA	0.08
INS.EFF_3DM_RAWNASA	0.08
LR.EFF_PM_RSME	0.07
TR.EFF_DM_RAWNASA	0.07
INS.EFF_3DM_RSME	0.07
TR.EFF_DM_RSME	0.05
LR.EFF_DM_RSME	0.05
INS.EFF_3DM_RAWNASA2	0.04

Table 4.10: Accuracy for the multinomial logistic regression classifiers grouped by efficiency score predictor

ciency scores provided the most accuracy with 19% and 15% respectively. The classifier built using INS.EFF_3DM_RAWNASA2 score produces the least accuracy with 4%. It is also interesting to note that the classifiers built using the TR.EFF_DM_RSME and LR.EFF_DM_RSME scores produced the second lowest accuracy results with 5% each.

Table 4.11 provides a breakdown of the accuracy scores for support vector machine classifiers using linear kernel built with the efficiency scores as the sole predictor variable to predict to predict the nodules to which an observation belongs. The linear kernel results were selected for discussion as it provided the best overall average of accuracy. Evaluation metrics for the other types of kernels can be found in Table B.4.

Predictor	**Accuracy**
LR.EFF_PM_RAWNASA	0.18
TR.EFF_PM_RAWNASA	0.15
LR.EFF_DM_RAWNASA	0.12
LR.EFF_LM_RAWNASA	0.11
LR.EFF_LM_RSME	0.11
TR.EFF_DM_RAWNASA	0.11
TR.EFF_LM_RSME	0.11
INS.EFF_3DM_RAWNASA2	0.09
INS.EFF_3DM_RSME2	0.09
LR.EFF_DM_RSME	0.09
LR.EFF_PM_RSME	0.09
TR.EFF_LM_RAWNASA	0.09
INS.EFF_3DM_RAWNASA	0.08
INS.EFF_3DM_RSME	0.08
TR.EFF_DM_RSME	0.07
TR.EFF_PM_RSME	0.05

Table 4.11: Accuracy for the support vector machine classifiers using linear kernel grouped by efficiency score predictor

It can be observed that the classifiers built using the LR.EFF_PM_RAWNASA and TR.EFF_PM_RAWNASA efficiency scores provided the most accuracy with 18% and 15% respectively. The accuracy scores produced by PM_RAWNASA score classifier models are consistent between the multinomial logistic regression (Table 4.10) and support vector machine (linear kernel) algorithms. On the other hand, classifier built using TR.EFF_PM_RSME score produces the least accuracy with 5%. This is 3% lower than the accuracy produced by the model built using a multinomial logistic regression algorithm.

4.2 Evaluation

Non-normal distribution of the variables could have resulted due to a number of reasons; outliers, insufficient data, and data collection method used. Outliers increase the variability within the dataset and cause the data to become skewed, which was observed in the dataset. The mean value is sensitive to outliers. Under normal circumstances, outliers should be removed and the data should be explored again. Detected outliers were not removed in the dataset, as the dataset was already small in size, especially in a few modules, where there were less than 20 samples. Removing the outliers would have meant that an important piece of information within a particular module may have gone unnoticed.

The presence of bimodal distributions would have potentially indicated the presence of two different groups within the dataset. However, none were observed. This suggests that the groups were more similar than expected and justifies the need for further exploration using statistical methods.

Concurrent validities between different models of efficiency should not follow a particular pattern. All the models are theoretically different to each other and should not behave like others. Across the models, the correlations should be unstable. The parabolic model takes more dynamics into account. Therefore, the assumption here was that there would be moderate correlation between them. However, as observed in Figure 4.7, all the resulting statistically significant correlations between the efficiency scores pairs of the different models of efficiency were large in size. It shows that the different models of efficiency could be measuring the same learning outcome with a large correlation, which is in support of the proposed hypothesis that it achieves moderate concurrent validity amongst the other models of efficiency.

Correlations for discriminant validity are meant to be low, if not neutral. Within models, the correlations should be consistent. None of the models of efficiency examined in this research demonstrated high discriminant validity, when comparing their training and learning efficiency scores. This is possibly due to the fact that both efficiency scores are calculated using the same performance measure, rather than using two different performance measures and mental effort or workload scores. This research collected repeated measurements taken on the same experimental unit (mental effort or workload) at two different time points (Pre-MCQ & Post-MCQ), but there was only one measure of MCQ scores. The correlation pairs of the parabolic model showed the least amount of correlation between the pairs, which is encouraging and supports the proposed hypothesis that a higher discriminant validity can be achieved when compared to the other models of efficiency.

Just because statistically significant differences were not observed on a majority of the Kruskal Wallis-tests, it should not be concluded that the two distributions do not differ. Those are only a comparison of the central tendency measure. The entire distribution needs to be considered to determine whether they differ or not. The outcomes of the shift function must also be considered. It is not a fair comparison when models which use absolute scales such as likelihood model & Parabolic model are compared to models which use relative scales such as deviational model and multidimensional model as the majority of the values for models which use relative scales will lie between +/- 3. This is not the case for models with absolute scales. They range from 0 to extensive positive values.

Another point to remember is that there are 2 efficiency scores (training & learning efficiencies) per score of mental effort / workload for each of the two-dimensional models (LM, PM & DM), where as there is only one such score for the multidimensional model (Instructional efficiency) because it uses two variables to calculate this efficiency. Therefore, it is not a fair comparison if these models are compared on the basis of the highest number of significant differences, as shown in Figure 4.12. Therefore, the results are presented for inclusion and completeness. It allows the researcher to view the broader picture. Parabolic model achieved the least amount of statistically significant differences. This does not support the proposed hypothesis that it is possible to achieve higher sensitivity.

Table 4.12 shows a breakdown of the number of statistically significant quantile difference per module. It is interesting to note that significant results were only shown in 11 of the 20 modules. Module 18 showed the highest number of differences amongst the modules with 41 quantiles out of 48 potential quantiles (16 scores x 3). Module 18 does not have the highest number of participants in the dataset and yet, it showed the best results by a margin of 29 quantiles, with the next best performing module showing only 12 quantile differences out of 48 potential quantiles.

Module	**Statistically Different Quantiles**
Module 18	41
Module 05	12
Module 08	11
Module 12	8
Module 13	7
Module 02	6
Module 04	4
Module 17	3
Module 06	2
Module 14	2
Module 16	1

Table 4.12: Modules with number of statistically significant quantile differences

Figure 4.16 shows the shift function plot and scatter plot of the LR.EFF_LM_RSME efficiency score for Module 1. The shift function plot shows zero statistically significant quantile difference. From examining the scatter plot in Figure 4.16, it can be seen that there are two outliers with efficiency scores above 63. Leaving those out, it can be observed that the range of efficiency scores seem to be relatively consistent between the control group (orange) and the experimental group (cyan). This is an indication of why there was no significant different quantiles identified by the shift function. The outliers from the control and experimental group seem to be in the third quantile, based on the efficiency scores and since efficiency score for the control group outliers is higher than the experimental group outlier, the shift function shows a slight difference for third quantile, although it is statistically not significant due to the confidence interval, which is represented as the orange vertical line, cross ”0” on the x-axis.

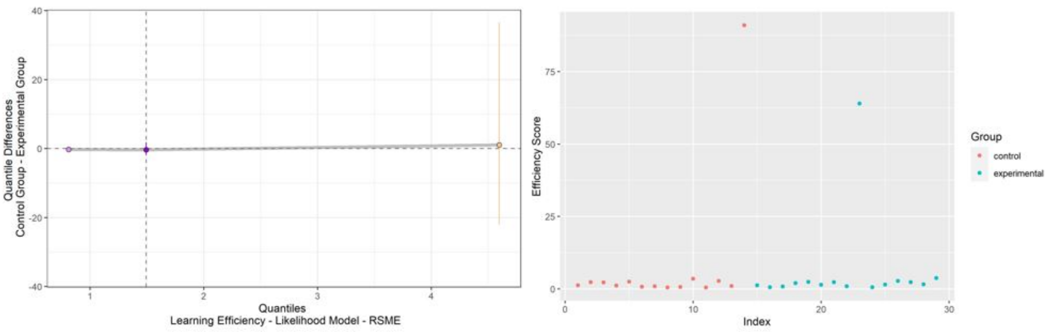

Figure 4.16: Comparison: Shift function plot and scatter plot of efficiency score - LR.EFF_LM_RSME - Module 1

Figure 4.17 shows the shift function plot and scatter plot of the LR.EFF_DM_RSME efficiency score for Module 2. The shift function plot shows one statistically significant difference at the first quantile. The quantile is in purple showing that the experimental group had the better efficiency. It can be observed in the scatter plot that there is an immediate difference in the range of efficiency scores between the control group and the experimental group. The control group has an approximate range between -2.3 and 1.7, whereas the experimental group has a range between -0.8 and 1.9 . It is easy to understand that there is a big difference between the efficiency scores of both groups at the first quantile. While there seems to be a difference between the two groups at the second and third quantiles, it is not statistically significant, represented by the confidence interval line.

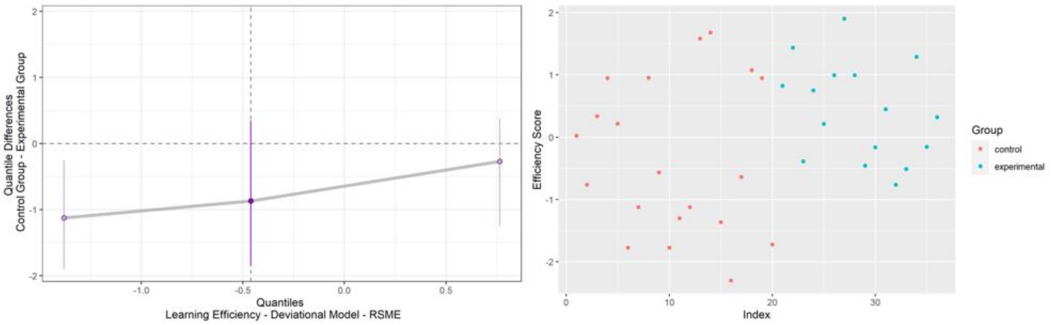

Figure 4.17: Comparison: Shift function plot and scatter plot of efficiency score - LR.EFF_DM_RSME - Module 2

Figure 4.18 shows the shift function plot and scatter plot of the LR.EFF_LM_RAWNASA efficiency score for Module 8. The shift function plot shows two statistically significant differences at the second and third quantiles. The quantiles are in orange showing that the control group had the better efficiency. It can be observed in the scatter plot that the control group has an approximate range between 0.9 and 8.2, whereas the experimental group has a range between 1.4 and 2.8. The control group has a much bigger spread of efficiency scores, whereas the experimental group has a smaller cluster. It is also important to note that the sample size for this module is quite small (16), which is less than the recommended amount as proposed by Wilcox, Erceg-Hurn, Clark, and Carlson (2014). While there seems to be a difference between the two groups at the first quantile, it is not statistically significant, represented by the confidence interval line.

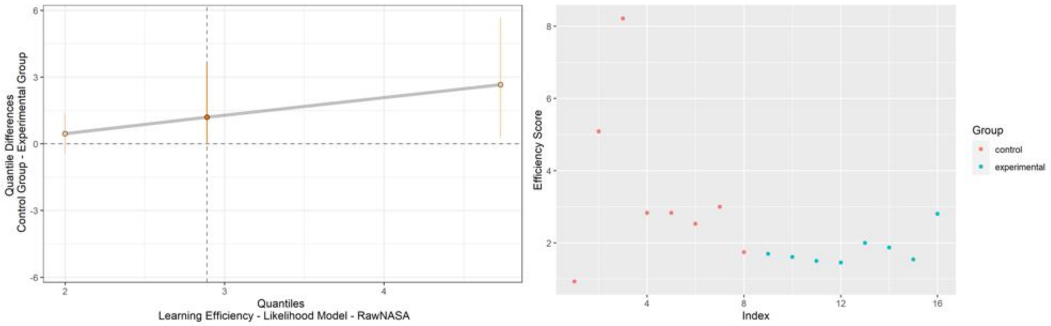

Figure 4.18: Comparison: Shift function plot and scatter plot of efficiency score - LR.EFF_LM_RAWNASA - Module 8

Figure 4.19 shows the shift function plot and scatter plot of the TR.EFF_LM_RSME efficiency score for Module 18. The shift function plot shows three statistically significant differences at all three quantiles. The quantiles are in purple showing that the experimental group had the better efficiency. It can be observed in the scatter plot that the control group has an approximate range between 0.5 and 3.3, whereas the experimental group has a range between 1.1 and 3.8. For the experimental group, the scatter plot shows three distinct clusters of efficiency scores, which is potentially indicative of the three quantiles for the group. however, it is not as apparent for the control group.

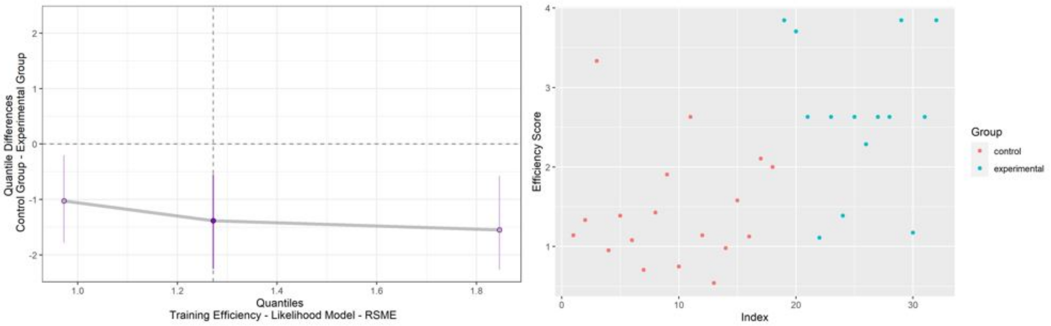

Figure 4.19: Comparison: Shift function plot and scatter plot of efficiency score - TR.EFF_LM_RSME - Module 18

The shift function provides much more information than the standard difference tests approach (Wilcox et al., 2014). Although the shift function is powerful, it also has its limitations. It can only be used with $\alpha = 0.5$ and it does not work well with tied values. The conclusions made from the quantile differences that the likelihood model efficiency scores differentiate between the two groups are tentative given the small sample size, which explains the large confidence intervals. The criteria on which the differences were accepted as statistically different was very tight in some cases and would be difficult to decide when viewed by the naked eye. There was an element of human judgement which decided which were accepted as statistically different and that must be taken into account for this analysis. Parabolic model achieved the lowest amount of statistically significant differences. This does not support the proposed hypothesis that it is possible to achieve higher sensitivity using this model.

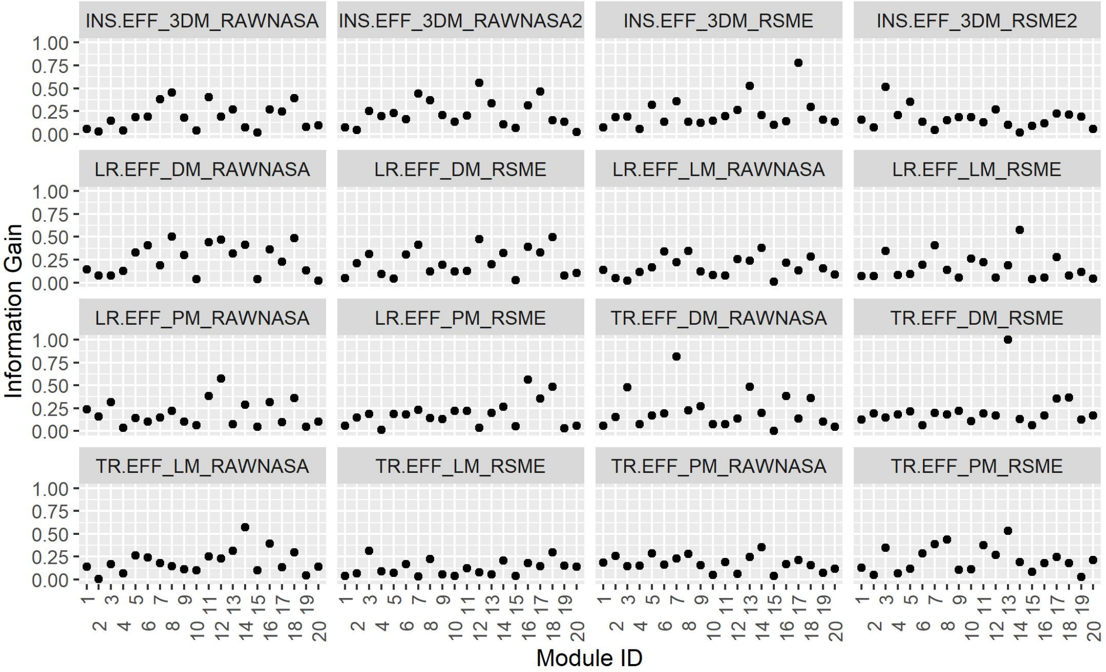

Figure 4.20: Scatter plots of the information gain units per modules grouped by efficiency score

Minimising the entropy will result in maximising the information gain (Murphy, 2012). However this is not something that can be achieved by data transformation of any kind. The entropy quality relies on the data collected. Having more samples that belong to a certain category will result in data that's more "pure" which in turn lowers entropy for that variable. For this research, it was important that a balanced dataset was achieved, and so entropy was not going to be minimised. Figure 4.20 shows the spread of I.G. units for each efficiency score across the 20 modules. by comparing the distribution of the I.G. units of each efficiency scores, it can be determined which score is more consistent. From examining the scatter plots, the efficiency score which seems to have the most consistent units of I.G is TR.EFF_LM_RSME across the modules. Based on the results shown in Table 4.7, it is expected that the classifiers built using LR.EFF_DM_RAWNASA as the predictor should produce the highest accuracy when classifying groups because it contained the most information in relation to the control v experimental group segmentation. Conversely, classifiers built using TR.EFF_LM_RSME as the predictor should produce the lowest accuracy when classifying groups. However, this was not the case. Classifier built with INS.EFF_3DM_RSME had the best predicting capability to classify between the groups correctly with accuracy and F-1 scores above 60% in both types of classifier models: binomial logistic regression and support vector machine - radial kernel.

Theoretically, one potential reason for achieving average accuracy scores when predicting the right group using the classifiers is because the overall dataset, which included information about all modules, was used to build classifiers instead of doing so per module. However as it has been mentioned before, the data available is insufficient to do it per module.

Imbalanced data is a common issue when it comes to classification problems, and it was encountered during this research. When groups are underrepresented,e.g, number of observations in each module, the class distribution starts skew. Balancing the classes within the "module" variable was required. Various strategies were considered to deal with the imbalanced data problem. Oversampling of the minority classes, under sampling of the majority classes and creating synthetic data were both considered. With only 455 observations, the option of under-sampling the majority classes was ruled out as useful information could be discarded. Oversampling the minority classes by replicating them to a constant degree. There is no information lost using this method, however this increased the likelihood of over fitting. Synthetic Minority Over-Sampling Technique (SMOTE) was considered to create synthetic data as proposed by Chawla, Bowyer, Hall, and Kegelmeter(2002) to induce more inferential statistics. SMOTE algorithm create similar samples from the minority class instead of repeating them. This techniques would have been ideal, however, while generating synthetic samples, the algorithm did not take into consideration than neighbouring samples can be from other classes. This introduced additional noise in the dataset by increasing the overlapping of classes. Since the data is synthetically created at random, it did not be reflect the conditions of the instructional design.

Classifiers built using efficiency scores based on the parabolic model achieved the lowest amount of accuracy and F-1 scores when predicting the groups. This does not support the proposed hypothesis that it is possible to achieve higher sensitivity. However, it produced the highest accuracy when predicting the modules. This is an encouraging find and it shows that the efficiency scores based on the parabolic model have some differentiating capabilities when there are multiple classes involved.

There is no clear reason observed as to why the parabolic model of efficiency was not able to better discriminate between the control and experimental groups clearly, particularly with the Known Groups validity where it performed the worst among all the models of efficiency. Evidence for this could be discovered once more research is carried out and further analysis is performed in the future.

4.3 Summary

The aim of this experiment was not only to find out the discriminating capability of the parabolic model, but also to find out to what extent it measures the same conceptual outcome as the other established models of efficiency.

The concurrent validity of the parabolic model was examined, compared against other two-dimensional models and evaluated. Concurrent validity was assessed by performing a correlation test between the training efficiency scores of all three 2-dimensional models in pairs and the learning efficiency scores of three 2-dimensional models in pairs. Concurrent validity is demonstrated if the efficiency scores from parabolic model correlates highly with the efficiency score from another model. The model performed better than expected and results show that the parabolic model has high concurrent validity among the the statistically significant correlations.

The discriminant validity of the parabolic model was examined, compared against other two-dimensional models and evaluated. Discriminant validity was assessed by performing a correlation test between the training and learning efficiency scores for all three 2-dimensional models for both measures of mental effort / workload. Discriminant validity is demonstrated, if the training efficiency score of a model has low correlation with its learning efficiency score. All the statistically significant correlations showed high correlation which does not demonstrate discriminant validity in general, however, the parabolic model resulted in the lowest average r_s among the three models under investigation. This shows that this model has a higher discriminant validity compared to the other two models.

Sensitivity of the parabolic model was examined, compared and evaluated. Sensitivity was assessed in three ways; Kruskal-Wallis test, Shift function and Classification. Each efficiency score was used as the sole basis for the tests / techniques to assess the sensitivity and determine each efficiency score's capability to discriminate between the control and experimental groups. The parabolic model was the poorest performer with the Kruskal-Wallis tests with the lowest amount of statistically significant differences across the 20 modules. The model also showed poor sensitivity when examined using the shift function. The classifiers built using efficiency scores based on the parabolic model showed average results compared to the other models when differentiating between the groups.

Results from the experiment shows partial evidence in favour of the proposed hypothesis. The parabolic model achieved moderate concurrent validity with an average r_s 0.64 and had a higher discriminant validity with the lowest average r_s amongst all three of the two-dimensional models using RAWNASA ($r_s = 0.71$) and RSME ($r_s = 0.62$) respectively. However, the model did not achieve higher sensitivity when compared to the other models. Based on the findings, there is no statistically significant evidence to reject the null hypothesis.

4.4 Strengths and Limitations of the Results

4.4.1 Strengths

This study has a specific focus on comparing the parabolic model of efficiency to other models. A focus which is notably new based on previous research in the field of education. The research findings will form the basis for further research in the future.

An interesting find from this study is the accuracy scores of the parabolic model based classifiers predicting the modules to which the observations belong with a small sample of data. Such classifiers demonstrated some discriminating capabilities to better identify the modules, which suggests that this model is somewhat sensitive to differences within the distributions.

The methods used in this experimentation for validity, sensitivity and classification, along with the metrics are broadly accepted in the field of science and education. Moreover, data collection for the experiment was conducted in real educational environment. Consequently, the collection of the data might have been affected by the noise which characterizes the participant groups, but is reflective of the third level educational set up.

The results obtained from using both variations of the multidimensional model of efficiency provide similar results of the sensitivity although difference was observed between them. The model with the modified formula yielded better results with the shift function and classification techniques, whereas the model with the original formula provided better results with the Kruskal-Wallis test.

The experimental framework explored is easily repeatable provided the right conditions are met and the results are reproducible if the same data is used, leading to future work based on this research.

4.4.2 Limitations

Detected outliers were not removed in the dataset, as the dataset was already small in size, especially in a few modules, where there were less than 20 samples. Removing the outliers would have meant that we may have missed out on an important piece of information within a particular module. This in turn may have increased the variability within the dataset which reduces the statistical power of the models. The decision was based on a trade-off between statistically significant results and having enough samples in the dataset to apply for the various techniques.

Kruskal-Wallis test has slightly lower power when compared to the parametric equivalent, ANOVA. This research collected repeated measurements taken on the same experimental unit at two different time points (Pre-MCQ & Post-MCQ). It was later realised, and possibly too late, that perhaps a Friedman test may have been more appropriate for determining the differences for such measurements.

The research lacked an adequate amount of data in general. Assumptions for the parametric tests were not met due to the limited dataset. This made it difficult to generalise the results. More statistically significant results could potentially have been achieved had there been a larger dataset. Theoretically, building a classifier model per module each with n $<$ 40 observations is not an ideal situation, as there is not enough data.

Classifying the modules was not a main focus of this experiment; it was conducted as supplement to the classification of the groups. Entropy scores and information gain were not calculated on that basis. Decision tree was not considered for classifier model, even though information gain was calculated because a decision tree normally requires multiple predictor variables in order to create various decision rules at various stages to predict the outcome. The experiment design specifies using a single variable as the predictor. Another reason is that decision tree algorithms do not have the same level of predictive accuracy as other approaches without the use of aggregation methods such as bagging and boosting, because a small change in the data can cause a large change in the final ”tree”.

Achieving statistically significant results is difficult while applying the shift function to data which contain n $<$ 20 observations. Although this function was applied to all the 20 modules, the results obtained must be examined carefully before coming to a conclusion.

Chapter 5

Conclusion

This chapter concludes the research by providing a summary of the work carried out, highlighting the contribution to the general body of research within instructional efficiency in third level education. Further areas of investigation and research will be addressed in order to potentially improve on the results found for future work.

5.1 Research Overview

Cognitive Load Theory is a widely known theory in educational psychology. It assumes that working memory can process only explicit instructions. Another method is the inquiry activity, under social constructivism theory, which is aimed at engaging learners by the use of focused communication focused on reaching an agreement and construct knowledge collaboratively. Research have been conducted in the past on teaching methodologies that aims at combining the traditional teaching method and a community of inquiry approach by extending the former with the latter and comparing its efficiency (experimental) versus the efficiency of traditional method alone (control).

Efficiency in learning and instruction is the capacity to achieve established goals with minimum expenditure of effort or resources. Efficiency is calculated based on the mental effort or workload exerted during a task and the performance outcome. Ideally, any activity conducted should be as efficient as possible. There have been various models of instructional proposed for use in the field of education. This research attempts to introduce a novel model of efficiency for comparison and assess whether it is suitable for application.

5.2 Problem Definition

It has emerged in the past that in education, the assumption / rationale that underpins efficiency is that low mental effort with high performance scores provides the best efficiency. By contrast, high mental effort with low performance provides the worst efficiency. Although the framework of optimal effort / mental workload is applied widely in other fields, it is not widely used when it comes to instructional efficiency. Another problem with the current models of efficiency is that either they are affected by variability of all the observations in the group or that they are sensitive to minor changes in the sample of observations. The parabolic model assumes that optimal workload and high performance provides the best efficiency and looks to address the concerns stated above.

The aim of this research is to evaluate the effectiveness of the parabolic model. This paper looked to introduce the novel model of instructional efficiency suited for education and evaluate the model's validity and sensitivity so that its credibility can be assessed. The model was compared with other state-of-the-art models of efficiency currently employed in third level education and the goal was to determine if the novel model better discriminates between participants of two distinct groups, control and experimental, based on their resulting efficiencies.

The comparison of groups typically involves the central tendency which are not robust. It assumes that distributions differ only in the central tendency. It is important to consider and compare entire distributions. This paper looked to include inferential statistics as well as descriptive statistics in order to evaluate the models and arrive at a conclusion.

It is vital to develop models of efficiency that are relevant to education. The study was limited to environments that only use the traditional method approach to learning so that the comparison can be justified. There were a number of issues with small sample sizes, typical in third level education which must be addressed where possible.

5.3 Design/Experimentation, Evaluation & Results

An empirical experiment was designed to test the hypothesis that it is possible to achieve higher sensitivity, discriminant validity and moderate concurrent validity using the parabolic model of efficiency when compared to the other state-of-the-art models such as likelihood model, deviational model and the multidimensional model. The experiment compared two instructional design conditions using the various models of efficiency. The data was collected from various modules in Technological University Dublin. Participants were divided into two groups and each group was allocated to a particular instructional design. Once the instructions were complete, the participants undertook a MCQ test and filled in questionnaires related to perceived mental effort both before the MCQ and after the MCQ. The collected data was analysed for normality, outliers, missing values etc. Then, the required mental workload measures (RAWNASA), standardised scores and efficiency scores (x 16) based on all models being compared were calculated.

The parabolic model was examined, compared and evaluated under different criteria: Validity and Sensitivity. Various test and techniques were used to asses the validity and sensitivity of the parabolic model and compare it to the other models. These include correlation tests, Kruskal-Wallis test, Shift function, Information gain and Classification. Modelling and evaluation criteria were specified at the design phase and the experiment was carried out as planned.

The results achieved in this empirical research showed that:

- Parabolic model of efficiency achieved high correlation with other models of efficiency which demonstrated concurrent validity.

- Parabolic model achieved high correlation between its training and learning efficiency. This was observed for all two-dimensional models of efficiency. The parabolic model achieved the lowest correlation among the other models, and therefore shows that this model has the higher discriminant validity.

- Parabolic model performed rather poorly when the known groups validity was assessed. This model showed the least amount of statistically significant results among all the models.

- Parabolic model performed poorly when the shift function was employed, similar to the known groups validity. The model showed the least amount of statistically significant quantile differences among all the models.

- Efficiency scores based on the parabolic model have moderate information gain when compared to efficiency scores based on other models. This wasn't the best performing model, but was not the worst either.

- Classifiers built using the efficiency scores based on the parabolic model provided moderate results when compared to the other efficiency scores when predicting the group (control v experimental)to which an observation belongs with accuracy ranging between 0.53 to 0.55. Interestingly, classifiers built using the efficiency scores using RAWNASA based on the parabolic model provided the best accuracy when compared to the other efficiency scores when predicting the module to which an observation belongs with accuracy ranging between 0.15 to 0.19.

The experiment experienced some issues such as small samples sizes and imbalanced dataset which were acknowledged. After evaluation of the results, it was concluded that there was there is no statistical evidence to reject the null hypothesis based on the findings, although there was partial evidence to support the proposed alternate hypothesis.

5.4 Contributions and Impact

This paper provided an insight into the application and comparison of a novel model of efficiency in the field of education. It sought to examine whether this novel model has better discriminating capabilities compared to the current state-of-the-art models, as well as examine its validity. It also provided an insight into issues highlighted such as lack of data, outliers, bias which future researchers may encounter on a similar research. To the best of the author's knowledge, no other piece of research examined the parabolic model of efficiency. This study offers a contribution to the use of shift function to determine the difference between groups in the educational context, and from this determining whether the results from different instructional designs actually provide any significant results.

Through the literature review, it provides an amalgamation of instructional design and multiple models of efficiency and explains how these concepts and methods are important to the human cognitive architecture. The research incorporates the concept of mental workload using both uni-dimensional and multidimensional measures. A significant advantage of the design framework is that it can be replicated and adapted in the future to expand on the research carried out. This study was based on instructional designs and model of efficiency which are important aspects of third level education. With the current situation in the world, different instructional designs could based on this design framework and the model of efficiency explored in this study could be assessed further.

Strengths and limitations were then highlighted with a view to understanding the process and to come up with recommended areas of future research. Further investigation and empirical research needs to be carried out to strengthen this contribution and confirm the potential of parabolic model of instructional efficiency as a novel method.

5.5 Future Work & Recommendations

Future work on instructional efficiency should focus on the findings from Chapter 4 and could concentrate on the following to improve the design of the experiment:

Collect and experiment using primary and secondary task measures, as well as physiological measures where possible and apply them with the models of efficiencies to investigate which model of efficiency is more sensitive to differences and discriminate better between the control and experimental groups. It could potentially remove any informational bias in the dataset, leading to more accurate results. This will help better evaluate the models and determine which model is suited for third level education. The application does not have to be confined to education. Other fields of application should be explored.

Further statistical tests for small sample size comparisons should be explored. Extend the use of shift function to more complex designs, to quantify interaction effects between various factors. Theoretically, it can be done. By performing a different analysis, we can better understand the different models of efficiency.

Another suggestion for future work would be to target bigger classes. A large sample size would assist with the research and open avenues to explore techniques which require large amount of data. This can be supplemented by designing classes with various levels of difficulty (easy/medium/hard). With a varying degree of difficulty and complexity, it will enable the researcher to gather more data, such as mental workload and performance, along the broader spectrum of scale which in turn would generate efficiency score along a bigger spectrum for better analysis of the models. Researchers should also consider streamlining the testing aspect of the experiment such as multi-choice questionnaire (MCQ) to collect performance measures. Researchers should ensure that the number of questions required for a test is consistent across the classes and that an ideal amount of questions is included. There is always a trade-off with the quantity of questions in a MCQ. The choice must be made by the researcher to determine the amount which is both meaningful and feasible to allow the participants to complete the task by the stipulated time. The time allowance will could also be a factor, whether it is applied during instruction or test. Another avenue that could be explored is the use of weighted answers for MCQ whereby each answer option would have a particular weighting assigned to it. This will also assist with collecting the performance data on the broader spectrum for better analysis.

References

Atkinson, R., & Shiffrin, R. (1968). Human memory: A proposed system and its control processes. In (Vol. 2, pp. 89–195). Academic Press. doi: 10.1016/S0079-7421(08)60422-3

Baddeley, A. (2001, 12). Is working memory still working? *The American psychologist*, *56*, 851–864. doi: 10.1027//1016-9040.7.2.85

Beckmann, J. F. (2010). Taming a beast of burden – on some issues with the conceptualisation and operationalisation of cognitive load. *Learning and Instruction*, *20*(3), 250–264. doi: https://doi.org/10.1016/j.learninstruc.2009.02.024

Bittner Jr., A. C., Byers, J. C., Hill, S. G., Zaklad, A. L., & Christ, R. E. (1989). Generic workload ratings of a mobile air defense system (los-f-h). *Proceedings of the Human Factors Society Annual Meeting*, *33*(20), 1476–1480. doi: 10.1177/154193128903302026

Carlson, K., & Herdman, A. (2012). Understanding the impact of convergent validity on research results. *Organizational Research Methods*, *15*, 17–32. doi: 10.1177/1094428110392383

Chandler, P., & Sweller, J. (1992). The split-attention effect as a factor in the design of instruction. *British Journal of Educational Psychology*, *62*, 233–246. doi: 10.1111/J.2044-8279.1992.TB01017.X

Chawla, N., Bowyer, K., Hall, L., & Kegelmeter, W. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi: 10.1613/jair.953

Cohen, J. (1988). Chapter 4.2 - the effect size index: q. In *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd edition ed.). Lawrence Erlbaum Associates.

Cohen, L., Manion, L., & Morrison, K. (2005). *Research methods in education.* RoutledgeFalmer.

De Jong, T. (2010). Cognitive load theory, educational research, and instructional design: some food for thought. *Instructional Science*, *38*(2), 105–134. doi: 10.1007/s11251-009-9110-0

Dewey, J. (2007). *Logic: The theory of inquiry.* Read Books.

Fischer, S., Lowe, R. K., & Schwan, S. (2008). Effects of presentation speed of a dynamic visualization on the understanding of a mechanical system. *Applied Cognitive Psychology*, *22*(8), 1126–1141. doi: 10.1002/acp.1426

Garrison, D. (2007, 01). Online community of inquiry review: Social, cognitive, and teaching presence issues. *Journal of Asynchronous Learning Networks*, *11*. doi: 10.24059/olj.v11i1.1737

George, D., & Mallery, P. (2010). *Spss for windows step by step: A simple guide and reference 18.0 update* (11th ed.). USA: Prentice Hall Press.

Gerjets, P., Scheiter, K., & Catrambone, R. (2006). Can learning from molar and modular worked examples be enhanced by providing instructional explanations and prompting self-explanations? *Learning and Instruction*, *16*(2), 104–121. doi: 10.1016/j.learninstruc.2006.02.007

Ghanbary Sartang, A., Ashnagar, E., M.and Habibi, & Sadeghi, S. (2016). Evaluation of rating scale mental effort (rsme) effectiveness for mental workload assessment in nurses. *Journal of Occupational Health and Epidemiology*, *5*(4), 211–217. doi: 10.18869/acadpub.johe.5.4.211

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. doi: 10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload* (Vol. 52, pp. 139–183). North-Holland. doi: 10.1016/S0166-4115(08)62386-9

Hendy, K. C., Hamilton, K. M., & Landry, L. N. (1993). Measuring subjective workload: When is one scale better than many? *Human Factors*, *35*(4), 579–601. doi: 10.1177/001872089303500401

Hoffman, B. (2012). Cognitive efficiency: A conceptual and methodological comparison. *Learning and Instruction*, *22*(2), 133–144. doi: 10.1016/j.learninstruc.2011.09.001

Hoffman, B., & Schraw, G. (2010). Conceptions of efficiency: Applications in learning and problem solving. *Educational Psychologist*, *45*(1), 1–14. doi: 10.1080/00461520903213618

Johnes, J., Silva, M. A., & Thanassoulis, E. (2017, 04). Efficiency in education. *Journal of the Operational Research Society*, *68*, 331–338. doi: 10.1057/s41274-016-0109-z

Jonassen, D. (2009). Reconciling a human cognitive architecture. In S. Tobias & T. M. Duffy (Eds.), *Constructivist instruction: Success or failure?* (1st ed., pp. 13–33). New York: Routledge. doi: 10.4324/9780203878842

Junior, A. C., Debruyne, C., Longo, L., & O'Sullivan, D. (2019). On the mental workload assessment of uplift mapping representations in linked data. In L. Longo & M. C. Leva (Eds.), *Human mental workload: Models and applications* (pp. 160–179). Cham: Springer International Publishing.

Kalyuga, S., & Sweller, J. (2005, 09). Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. *Educational Technology Research and Development*, *53*, 83–93. doi: 10.1007/BF02504800

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support vector machines in r. *Journal of Statistical Software*, *15*, 1–28. doi: 10.18637/jss.v015.i09

Kester, L., Lehnen, C., Van Gerven, P. W., & Kirschner, P. A. (2006). Just-in-time, schematic supportive information presentation during cognitive skill acquisition. *Computers in Human Behavior*, *22*(1), 93–112. doi: 10.1016/j.chb.2005.01.008

Kirschner, P., Sweller, J., & Clark, R. (2006, 06). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*, 75–86. doi: 10.1207/s15326985ep4102_1

Kirschner, P. A., Sweller, J., Kirschner, F., & Zambrano R., J. (2018, 06). From cognitive load theory to collaborative cognitive load theory. *International Journal of Computer-Supported Collaborative Learning*, *13*, 213–233. doi: 10.1007/s11412-018-9277-y

Krabbe, P. F. (2017). Chapter 7 - validity. In *The measurement of health and health status* (pp. 113–134). San Diego: Academic Press. doi: https://doi.org/10.1016/B978-0-12-801504-9.00007-6

Liu, Y., & Wickens, C. D. (1994). Mental workload and cognitive task automaticity: an evaluation of subjective and time estimation metrics. *Ergonomics*, *37*(11), 1843–1854. doi: 10.1080/00140139408964953

Longo, L. (2011). Human-computer interaction and human mental workload: Assessing cognitive engagement in the world wide web. In *Ifip conference on human-computer interaction* (pp. 402–405).

Longo, L. (2012). Formalising human mental workload as non-monotonic concept for adaptive and personalised web-design. In *International conference on user modeling, adaptation, and personalization* (pp. 369–373).

Longo, L. (2014). *Formalising human mental workload as a defeasible computational concept* (Unpublished doctoral dissertation). Trinity College.

Longo, L. (2015, 03). A defeasible reasoning framework for human mental workload representation and assessment. *Behaviour and Information Technology*, *34*, 758–786. doi: 10.1080/0144929X.2015.1015166

Longo, L. (2016, 06). Mental workload in medicine: Foundations, applications, open problems, challenges and future perspectives. In (pp. 106–111). doi: 10.1109/CBMS.2016.36

Longo, L. (2018). On the reliability, validity and sensitivity of three mental workload assessment techniques for the evaluation of instructional designs: A case study in a third-level course. In *Proceedings of the 10th international conference on computer supported education, CSEDU 2018, funchal, madeira, portugal, march 15-17, 2018, volume 2.* (pp. 166–178). doi: 10.5220/0006801801660178

Longo, L., & Leva, M. C. (2017). *Human mental workload: Models and applications: First international symposium, h-workload 2017, dublin, ireland, june 28-30, 2017, revised selected papers* (Vol. 726). Springer.

Longo, L., & Orru, G. (2019). An evaluation of the reliability, validity and sensitivity of three human mental workload measures under different instructional conditions in third-level education. In *Proceedings of the 10th international conference on computer supported education: Csedu* (pp. 384–413). Springer, Cham. doi: 10.1007/978-3-030-21151-6_19

Longo, L., & Orrú, G. (2020). Evaluating instructional designs with mental workload assessments in university classrooms. *Behaviour & Information Technology*, *0*(0), 1-31. Retrieved from `https://doi.org/10.1080/0144929X.2020.1864019` doi: 10.1080/0144929X.2020.1864019

Miller, G. A. (1956). The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review*, *63 2*, 81–97. doi: 10.1037/h0043158

Moustafa, K., & Longo, L. (2019, 02). Analysing the impact of machine learning to model subjective mental workload: A case study in third-level education. In (pp. 92–111). doi: 10.1007/978-3-030-14273-5_6

Murphy, K. P. (2012). Entropy. In *Machine learning: A probabilistic perspective.* The MIT Press.

Nachreiner, F. (1995). Standards for ergonomics principles relating to the design of work systems and to mental workload. *Applied Ergonomics*, *26*(4), 259–263. (Ergonomics and International Standards) doi: 10.1016/0003-6870(95)00029-C

Orru, G., Gobbo, F., O'Sullivan, D., & Longo, L. (2018). An investigation of the impact of a social constructivist teaching approach, based on trigger questions, through measures of mental workload and efficiency. *Proceedings of the 10th International Conference on Computer Supported Education (CSEDU 2018)*, *2*, 292–302. doi: 10.5220/0006790702920302

Orru, G., & Longo, L. (2019a). Direct instruction and its extension with a community of inquiry: A comparison of mental workload, performance and efficiency. In *Proceedings of the 11th international conference on computer supported education, CSEDU 2019, heraklion, crete, greece, may 2-4, 2019, volume 1.* (pp. 436–444). Retrieved from `https://doi.org/10.5220/0007757204360444` doi: 10.5220/0007757204360444

Orru, G., & Longo, L. (2019b). The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and germane loads: A review. In L. Longo & M. C. Leva (Eds.), *Human mental workload: Models and applications* (pp. 23–48). Cham: Springer International Publishing.

Orru, G., & Longo, L. (2020). Direct and constructivist instructional design: A comparison of efficiency using mental workload and task performance. In L. Longo & M. C. Leva (Eds.), *Human mental workload: Models and applications* (pp. 99–123). Cham: Springer International Publishing.

Paas, F. (1992, 12). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology*, *84*, 429–434. doi: 10.1037/0022-0663.84.4.429

Paas, F., & Van Merrienboer, J. J. (1993). The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *35*(4), 737–743. doi: 10.1177/001872089303500412

Paas, F., Van Merrienboer, J. J. G., & Adam, J. (1994, 09). Measurement of cognitive load in instructional research. *Perceptual and motor skills*, *79*, 419–430. doi: 10.2466/pms.1994.79.1.419

Peterson, P. L. (1979, 10). Direct instruction: Effective for what and for whom? *Educational Leadership*, *37*, 46–48.

Plass, J. L., Moreno, R., & Brünken, R. (2010). *Cognitive load theory* (1st ed.). New York: Cambridge University Press.

Pollock, E., Chandler, P., & Sweller, J. (2002). Assimilating complex information. *Learning and Instruction*, *12*(1), 61–86. doi: 10.1016/S0959-4752(01)00016-0

Reed, S. K. (2012). Human cognitive architecture. In *Encyclopedia of the sciences of learning* (pp. 1452–1455). Boston, MA: Springer US. doi: 10.1007/978-1-4419-1428-6_328

Rizzo, L., Dondio, P., Delany, S., & Longo, L. (2016, 09). Modeling mental workload via rule-based expert system: A comparison with nasa-tlx and workload profile. In (Vol. 475, pp. 215–229). doi: 10.1007/978-3-319-44944-9_19

Rizzo, L. M., & Longo, L. (2017). Representing and inferring mental workload via defeasible reasoning: a comparison with the nasa task load index and the workload profile. In *Proceedings of the 1st workshop on advances in argumentation in artificial intelligence ai3ai-ia.*

Rubio Valdehita, S., Ramiro, E., García, J., & Puente, J. (2004, 01). Evaluation of subjective mental workload: a comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology*, *53*, 61 – 86. doi: 10.1111/j.1464-0597.2004.00161.x

Simina, V. K. (2012). Socio-constructivist models of learning. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 3128–3131). Boston, MA: Springer US. doi: 10.1007/978-1-4419-1428-6_882

Slocum, T. A. (2004). Chapter 6 - direct instruction: The big ideas. In D. J. Moran & R. W. Malott (Eds.), *Evidence-based educational methods* (pp. 81–94). San Diego: Academic Press. doi: 10.1016/B978-012506041-7/50007-3

Smith, P. C., & Street, A. (2005). Measuring the efficiency of public services: The limits of analysis. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *168*(2), 401–417. doi: 10.2307/3559969

Sweller, J. (2003, 12). Evolution of human cognitive architecture. In (Vol. 43, pp. 215–266). doi: 10.1016/S0079-7421(03)01015-6

Sweller, J. (2009). What human cognitive architecture tells us about constructivism. In S. Tobias & T. M. Duffy (Eds.), *Constructivist instruction: Success or failure?* (1st ed., pp. 127–143). New York: Routledge. doi: 10.4324/9780203878842

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*, 123–138. doi: 10.1007/s10648-010-9128-5

Sweller, J. (2011). Chapter two - cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), (Vol. 55, pp. 37–76). Academic Press. doi: https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Sweller, J. (2016). Cognitive load theory, evolutionary educational psychology, and instructional design. In D. C. Geary & D. B. Berch (Eds.), *Evolutionary perspectives on child development and education* (pp. 291–306). Cham: Springer International Publishing. doi: 10.1007/978-3-319-29986-0_12

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load as a factor in the structuring of technical material. *Journal of Experimental Psychology: General*, *119*, 176–192. doi: 10.1037/0096-3445.119.2.176

Sweller, J., Van Merrienboer, J. J. G., & Paas, F. (1998, 09). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*, 251–296. doi: 10.1023/a:1022193728205

Sweller, J., Van Merrienboer, J. J. G., & Paas, F. (2019, 06). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*, 261–292. doi: 10.1007/s10648-019-09465-5

Tindall-Ford, S., Chandler, P., & Sweller, J. (1997). When two sensory modes are better than one. *Journal of Experimental Psychology: Applied*, *3*, 257–287. doi: 10.1037/1076-898X.3.4.257

Tuovinen, J. E., & Paas, F. (2004). Exploring multidimensional approaches to the efficiency of instructional conditions. *Instructional Science*, *32*, 133–152. doi: 10.1023/B:TRUC.0000021813.24669.62

van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist*, *43*(1), 16–26. doi: 10.1080/00461520701756248

Van Merrienboer, J. J. G., & Sweller, J. (2005, 06). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*, 147–177. doi: 10.1007/s10648-005-3951-0

Virues-Ortega, J., Montaño-Fidalgo, M., Froján-Parga, M. X., & Calero-Elvira, A. (2011). Descriptive analysis of the verbal behavior of a therapist: A known-group validity analysis of the putative behavioral functions involved in clinical interaction. *Behavior Therapy*, *42*(4), 547–559. doi: https://doi.org/10.1016/j.beth.2010.12.004

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing.* Academic Press.

Wilcox, R. R., Erceg-Hurn, D. M., Clark, F., & Carlson, M. (2014). Comparing two independent groups via the lower and upper quantiles. *Journal of Statistical Computation and Simulation*, *84*(7), 1543–1551. doi: 10.1080/00949655.2012.754026

Wilcox, R. R., & Keselman, H. J. (2003). Modem robust data analysis methods: Measures of central tendency. *Psychological Methods*, *8*(3), 254–274. doi: 10.1037/1082-989X.8.3.254

Zijlstra, F., & Doorn, L. (1985, 01). The construction of a scale to measure perceived effort. *Department of Philosophy and Social Sciences*, 53.

Zijlstra, F. R. H. (1993). *Efficiency in work behavior: A design approach for modern tools* (Doctoral dissertation). Retrieved from `https://www.researchgate.net/\publication/27344359`

Appendix A

Original Dataset

Table A.1: Details of the original data collected

Variable	**Description**	**Type**	**Range / Values**
MCQ Score	Performance Score	Integer	13 - 100
Module	Name of module	Categorical	Various
Module ID	Module ID	Categorical	1 - 20
Date	Date the module was held	Date	07/02/2019 to 19/02/2020
Group	Group name	Categorical	control , experimental
PRE.Knowledge	Amount of knowledge the individual has in relation to the task. (Before taking MCQ)	Integer	1 - 20
PRE.Motivation	How much the individual is motivated to perform the task. (Before taking MCQ)	Integer	1 - 20

Variable	Description	Type	Range / Values
PRE.Effort	Amount of hard-work required to accomplish the task. (Before taking MCQ)	Integer	1 - 20
PRE.Frustration	Amount of emotional drainage and irritation. (Before taking MCQ)	Integer	1 - 20
PRE.Mental	Amount of mental activity required while performing the task (Before taking MCQ)	Integer	1 - 20
PRE.Performance	Amount of success in reaching the goal. (Before taking MCQ)	Integer	1 - 20
PRE.Physical	Amount of physical activity required while performing the task. (Before taking MCQ)	Integer	1 - 20
PRE.Temporal	Amount of time pressure felt while performing the task (Before taking MCQ)	Integer	1 - 20
PRE.RSME	Perceived Mental Effort rating (Before taking MCQ)	Integer	1 - 150
POST.Knowledge	Amount of knowledge the individual has in relation to the task. (After taking MCQ)	Integer	1 - 20

Variable	Description	Type	Range / Values
POST.Motivation	How much the individual is motivated to perform the task. (After taking MCQ)	Integer	1 - 20
POST.Effort	Amount of hard-work required to accomplish the task. (After taking MCQ)	Integer	1 - 20
POST.Frustration	Amount of emotional drainage and irritation. (After taking MCQ)	Integer	1 - 20
POST.Mental	Amount of mental activity required while performing the task (After taking MCQ)	Integer	1 - 20
POST.Performance	Amount of success in reaching the goal. (After taking MCQ)	Integer	1 - 20
POST.Physical	Amount of physical activity required while performing the task. (After taking MCQ)	Integer	1 - 20
POST.Temporal	Amount of time pressure felt while performing the task (After taking MCQ)	Integer	1 - 20
POST.RSME	Perceived Mental Effort rating (After taking MCQ)	Integer	1 - 150

ID	ID	Module	control	experimental	Total
1	A	Research Methods	14	15	29
2	B	Research Hypothesis	20	16	36
3	C	Visualising Geo Spatial Data	5	7	12
4	D	Operating Systems	20	18	38
5	E	Problem Solving	14	11	25
6	F	Data Mining	10	9	19
7	G	Literature Review	7	8	15
8	H	Research Hypothesis	8	8	16
9	I	Strings	10	12	22
10	J	Program Design	15	15	30
11	K	Machine Learning	5	8	13
12	L	Image Processing	7	9	16
13	M	Research Methods	8	9	17
14	N	Statistics	6	7	13
15	O	IT Forensics	19	14	33
16	P	Literature Comprehension	7	9	16
17	Q	Virtual Memory	8	7	15
18	R	Research Hypothesis	18	14	32
19	S	Literature Review	16	15	31
20	T	Operating Systems	14	13	27

Table A.2: Details of the 20 modules with a breakdown of participants in control and experimental group

Appendix B

Classification Results

Predictor	Accuracy	Precision	Recall	F-1 Score
INS.EFF_3DM_RAWNASA	0.57	0.43	0.56	0.48
INS.EFF_3DM_RAWNASA2	0.47	1.00	0.47	0.64
INS.EFF_3DM_RSME	0.62	0.63	0.59	0.61
INS.EFF_3DM_RSME2	0.47	0.23	0.40	0.29
LR.EFF_DM_RAWNASA	0.57	0.54	0.54	0.54
LR.EFF_DM_RSME	0.62	0.66	0.59	0.62
LR.EFF_PM_RAWNASA	0.47	1.00	0.47	0.64
LR.EFF_PM_RSME	0.47	1.00	0.47	0.64
LR.EFF_LM_RAWNASA	0.39	0.77	0.42	0.55
LR.EFF_LM_RSME	0.43	0.89	0.45	0.60
TR.EFF_DM_RAWNASA	0.46	0.23	0.38	0.29
TR.EFF_DM_RSME	0.58	0.57	0.56	0.56
TR.EFF_PM_RAWNASA	0.50	0.46	0.47	0.46
TR.EFF_PM_RSME	0.47	1.00	0.47	0.64
TR.EFF_LM_RAWNASA	0.46	0.97	0.47	0.63
TR.EFF_LM_RSME	0.45	0.91	0.46	0.61

Table B.1: Classification model metrics - Classifying group - Linear SVM kernel

Predictor	**Accuracy**	**Precision**	**Recall**	**F-1 Score**
INS.EFF_3DM_RAWNASA	0.51	0.03	0.33	0.05
INS.EFF_3DM_RAWNASA2	0.50	1.00	0.49	0.65
INS.EFF_3DM_RSME	0.61	0.89	0.55	0.68
INS.EFF_3DM_RSME2	0.46	0.03	0.14	0.05
LR.EFF_DM_RAWNASA	0.49	0.03	0.20	0.05
LR.EFF_DM_RSME	0.53	0.14	0.50	0.22
LR.EFF_PM_RAWNASA	0.53	0.54	0.50	0.51
LR.EFF_PM_RSME	0.59	0.26	0.69	0.38
LR.EFF_LM_RAWNASA	0.42	0.89	0.44	0.59
LR.EFF_LM_RSME	0.45	0.91	0.46	0.61
TR.EFF_DM_RAWNASA	0.45	0.23	0.36	0.28
TR.EFF_DM_RSME	0.54	0.54	0.51	0.53
TR.EFF_PM_RAWNASA	0.53	0.26	0.50	0.34
TR.EFF_PM_RSME	0.55	0.43	0.54	0.48
TR.EFF_LM_RAWNASA	0.47	1.00	0.47	0.64
TR.EFF_LM_RSME	0.45	0.94	0.46	0.62

Table B.2: Classification model metrics - Classifying group - Polynomial SVM kernel

Predictor	Accuracy	Precision	Recall	F-1 Score
INS.EFF_3DM_RAWNASA	0.54	0.60	0.51	0.55
INS.EFF_3DM_RAWNASA2	0.49	0.31	0.44	0.37
INS.EFF_3DM_RSME	0.55	0.37	0.54	0.44
INS.EFF_3DM_RSME2	0.62	0.43	0.65	0.52
LR.EFF_DM_RAWNASA	0.62	0.63	0.59	0.61
LR.EFF_DM_RSME	0.53	0.54	0.50	0.52
LR.EFF_PM_RAWNASA	0.45	0.37	0.41	0.39
LR.EFF_PM_RSME	0.53	0.17	0.50	0.26
LR.EFF_LM_RAWNASA	0.59	0.54	0.58	0.56
LR.EFF_LM_RSME	0.54	0.20	0.54	0.29
TR.EFF_DM_RAWNASA	0.49	0.29	0.43	0.34
TR.EFF_DM_RSME	0.46	0.31	0.41	0.35
TR.EFF_PM_RAWNASA	0.54	0.57	0.51	0.54
TR.EFF_PM_RSME	0.45	0.20	0.35	0.25
TR.EFF_LM_RAWNASA	0.55	0.23	0.57	0.33
TR.EFF_LM_RSME	0.55	0.09	0.75	0.15

Table B.3: Classification model metrics - Classifying group - Sigmoid SVM kernel

	Kernels		
Predictor	**Polynomial**	**Radial**	**Sigmoid**
INS.EFF_3DM_RAWNASA	0.09	0.08	0.08
INS.EFF_3DM_RAWNASA2	0.04	0.07	0.12
INS.EFF_3DM_RSME	0.07	0.07	0.05
INS.EFF_3DM_RSME2	0.07	0.07	0.11
LR.EFF_DM_RAWNASA	0.07	0.09	0.08
LR.EFF_DM_RSME	0.11	0.12	0.07
LR.EFF_PM_RAWNASA	0.16	0.14	0.14
LR.EFF_PM_RSME	0.09	0.09	0.09
LR.EFF_LM_RAWNASA	0.14	0.15	0.15
LR.EFF_LM_RSME	0.08	0.11	0.07
TR.EFF_DM_RAWNASA	0.11	0.09	0.07
TR.EFF_DM_RSME	0.09	0.11	0.09
TR.EFF_PM_RAWNASA	0.14	0.12	0.15
TR.EFF_PM_RSME	0.07	0.05	0.11
TR.EFF_LM_RAWNASA	0.11	0.12	0.14
TR.EFF_LM_RSME	0.07	0.07	0.07

Table B.4: Classification model accuracy - Classifying modules - Other SVM kernels