2018-1

An Application of Natural Language Processing for Triangulation of Cognitive Load Assessments in Third Level Education

Luis Alfredo Contreras

Technological University Dublin

Recommended Citation

Contreras, L. A. (2018) *An application of Natural Language Processing for triangulation of Cognitive Load Assessments in Third Level Education*. Masters dissertation, DIT, 2018.

An application of Natural Language Processing for triangulation of Cognitive Load Assessments in Third Level Education

Luis Alfredo Contreras

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Stream)

Date: January 2018

Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Stream), 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 Dublin Institute of Technology 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 Institutes guidelines for ethics in research.

Signed:

Date:

Abstract

Work has been done to measure Mental Workload based on applications mainly related to ergonomics, human factors, and Machine Learning. The influence of Machine Learning is a reflection of an increased use of new technologies applied to areas conventionally dominated by theoretical approaches. However, collaboration between MWL and Natural Language Processing techniques seems to happen rarely. In this sense, the objective of this research is to make use of Natural Languages Processing techniques to contribute to the analysis of the relationship between Mental Workload subjective measures and Relative Frequency Ratios of keywords gathered during pre-tasks and post-tasks of MWL activities in third-level sessions under different topics and instructional designs. This research employs secondary, empirical and inductive methods to investigate Cognitive Load theory, instructional designs, Mental Workload foundations and measures and Natural Language Process Techniques. Then, NASA-TLX, Workload Profile and Relative Frequency Ratios are calculated. Finally, the relationship between NASA-TLX and Workload Profile and Relative Frequency Ratios is analysed using parametric and non-parametric statistical techniques. Results show that the relationship between Mental Workload and Relative Frequency Ratios of keywords, is only medium correlated, or not correlated at all. Furthermore, it has been found out that instructional designs based on the process of hearing and seeing, and the interaction between participants, can overcome other approaches such as those that make use of videos supported with images and text, or of a lecturer's speech supported with slides.

Keywords: Mental Workload, Natural Language Processing, Instructional Design

Acknowledgements

After a period of intense learning, it is time to thank those who believed in me and supported me throughout this process.

I would like to thank my supervisor **Dr Luca Longo**, who cared so much about my work and guided and advised me without hesitation and with enthusiasm and professionalism.

I am also grateful to my employer, **Brightflag**, and particularly **Ian** and **Alex**, for allowing me to put my studies into practice. I have been enriched by this experience.

Finally, I must express my gratitude to those who have kept me going when times were tough. Leaving my country and family has been unthinkably difficult but I have some amazing angels who have given me, in different ways, a second chance for a new beginning. There are no words to sufficiently thank **Dr Sara Jane Delany**, **John McGrath**, **Michael Dineen** and last but most importantly **Dr Joe McGrath** who has been there from day one.

In honour of my Mum and Dad.

Contents

List of Figures

List of Tables

List of Acronyms

MWL	Mental Workload
NASA-TLX	NASA Task Load Index
WP	Workload Profile
NLP	Natural Language Processing
WUP	Wu and Palmer (1994) similarity method
FT	Relative frequency of a word in a core text
FC	Relative frequency of a word in a contrastive text

Chapter 1

Introduction

1.1 Background

Mental workload (MWL) is the amount of cognitive work that an individual requires to complete a task over time (Longo, 2016). It is a concept that is invoked when the complexity of that task needs to be measured, considering the interaction between factors related to the person, circumstances and the requirements of that activity (Hart, 2006; Longo & Barrett, 2010; Longo, 2015b). It has applications in ergonomics, human factors, computer systems and it is increasingly analysed in recent years with the collaboration of new technologies such as Machine Learning. Xie and Salvendy (2000) suggest that mental overload and underload have a negative effect on performance. Furthermore, Longo (2016) refers to mental underload as the stage when individuals may feel frustrated or annoyed whereas mental overload is related to the stage when an individual is confused, decreasing performance and increasing possible mistakes. In this sense, from a learning point of view, there are factors that may affect working memory load and such factors depend on the manner in which a material is presented or its intrinsic nature (Paas, Renkl, & Sweller, 2003). Thus, to achieve knowledge, the way in which information is presented and the learning activities required should be considered when planning instructional designs (Van Merrienboer & Sweller, 2005).

1.2 Research Problem

Work has been done to measure Mental Workload under instructional designs to enhance learning based on applications mainly related to ergonomics, human factors, and Machine Learning. The use of Machine Learning reflects the increased application of new technologies to areas conventionally dominated by theoretical approaches. However, collaboration between MWL and Natural Language Processing techniques seems to rarely happen. Boundaries of research of related works can be extended based on framing scholarships to contribute to knowledge in fields already studied. Although research has been done to analyse MWL, the need for more studies seems evident due to the complexity of measuring it. In this sense, the increased application of new technologies, such as Machine Learning, to MWL open up the question as to whether Natural Language Processing techniques can also make a contribution.

Furthermore, instructional designs should be planned to enhance learning. In this sense, if it is well planned, considering the way in which the information is presented and the learning activities required, it will have a positive impact on individuals learning, and it might contribute to achieving an optimal Mental workload. Thus, the content of an instrument design translated into text, and keywords of activity given under that instrument design, could be processed using NLP techniques and then analysed to measure how those keywords are related to the Mental Workload in terms of learning.

1.3 Research Methodologies

This experiment is considered as secondary, empirical and inductive research. It is a secondary analysis because the data was obtained from external sources. Namely, the sources were a dataset built by Dr Luca Longo from Dublin Institute of Technology and an open source called WordNet lexical database. It is empirical and inductive because this research is direct and measurable, establishing the inductive basis for future work based on the analysis between subjective measures of Mental Workload, instructional designs and the use of Natural Language Processing techniques to text related to topics taught by Dr Luca Longo.

1.4 Scope and Limitations

This research will use data gathered between 2015 until 2017, from Mental workload experiments that were conducted by Dr Luca Longo during third-level classes at Dublin Institute of Technology (DIT). During those experiments, different topics were taught using instructional designs where the NASA-TLX and Workload Profile questionnaires were used to collect factors and weights necessary to measure MWL. Along with the questionnaires, keywords were also collected from the participants during pre-tasks and post-tasks. Under those circumstances, the size of the data collected from those sessions is considered a limitation because the datasets contain 105 and 120 records respectively and they could not achieve statistically significant results when conducting the experiments.

1.5 Document Outline

This research involves five chapters, namely: Literature review and related work; Experiment design and methodology; Implementation and results; Evaluation; and Conclusion. A brief overview of their contents is presented as follows:

1. **Chapter 2** provides a literature review, related works and gaps in the fundamental fields of study which are necessary to formulate the research question. The fundamental fields are Instructional Design, Mental Workload and Natural Language Processing. Firstly, the Instructional Design section is presented. It begins with an analysis of Cognitive Load Theory, its definition and purpose, characteristics, and the relationship between it and instructional designs. Then, it covers types of instructions. The Mental Workload section analyses its foundations. Then, the main categories of the Mental Workload measures are presented followed by the subjective measures NASA Task Load Index and Workload Profile. Finally, the related work and summary sections present the work that has been done based on Instructional Designs, Mental Workload and Natural Language Processing and the gaps that motivate the formulation of the research

question.

2. **Chapter 3** presents a definition of the hypotheses necessary to answer the research question. It also involves software selection, data understanding, data preparation, model design, evaluation and hypotheses testing and strengths and limitations of the design approach.

3. **Chapter 4** provides the results of the performed data understanding, data preparation and modelling of the designed research. It begins with the process of inspecting the data used in this research. Its aim is to identify data quality problems and to discover insights. This involves a number of approaches to conduct the analysis. The data preparation involves the generation and reduction of features and the data processing of the corpus. Thus, data quality problems are handled, along with the assessment of normality of the variables. Finally, the modelling part is presented and the analysis of normality, linearity and homoscedasticity between variables is conducted.

4. **Chapter 5** involves the hypotheses testing and the reflection of strengths and limitations of findings based on the analyses performed during the previous sections.

5. **Chapter 6** presents the conclusion of this thesis, which involves: a research overview; problem definition; design experimentation, evaluation and results; contribution and impact; and future work recommendations.

Chapter 2

Literature review and related work

This chapter provides the literature review, related work and gaps in the fundamentals fields of study which are necessary to formulate the research question (see figure 2.1). These fundamental fields are Instructional Design, Mental Workload, and Natural Language Processing.

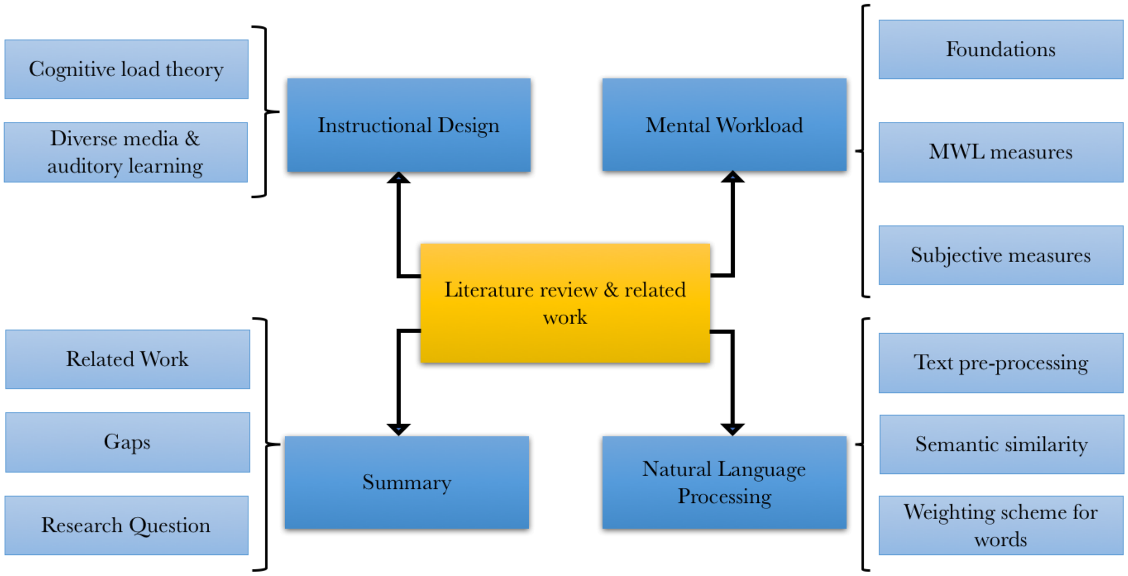


Figure 2.1: The Literature review process and related work which involves an overview of the fundamentals necessary to the formulation of the research question.

Firstly, the Instructional Design section is presented, which begins with an analysis of Cognitive Load Theory, its development, its definition and purpose, characteristics, and the relationship between it and instructional designs. Then, it covers types of instructions, namely diverse media and auditory learning, their importance and their related approaches.

The Mental Workload section outlines its foundations which describes the concept, applications and the negative impact of mental overload and mental underload in performance. Then, the main categories of the Mental Workload measures are presented followed by the subjective measures NASA Task Load Index and Workload Profile.

The Natural Language Processing section begins with the factors which have influenced the development of Natural Language Processing during the last ten years. Then, its definition is presented and the different procedures and applications commonly used are described. Also, it contextualises the techniques related to the approaches and forms, and the research question. Specifically, it contextualises techniques for text preprocessing, similarity measures and weighting scheme for words.

Finally, the related work and summary sections present the work that has been done based on Instructional Designs, Mental Workload and Natural Language Processing and the gaps that motivate the formulation of the research question.

2.1 Instructional Design

2.1.1 Cognitive Load Theory

The literature indicates that work in Cognitive load theory has been increasing since it was developed in 1980 (Sweller, Van Merrienboer, & Paas, 1998; Paas et al., 2003; Van Merrienboer & Sweller, 2005; Paas, van Gog, & Sweller, 2010; Leppink, Paas, Van der Vleuten, Van Gog, & Van Merriënboer, 2013; Paas & Ayres, 2014; Kalyuga & Singh, 2016; Schilling, 2017; Costley & Lange, 2017). Cognitive load theory studies the ease with which information may be processed in working memory (Sweller et al., 1998) and examines ways in which the working memory is used to transfer information into the long-term memory (Costley & Lange, 2017). Paas et al. (2003) suggest that there are factors that may affect working memory load and that such factors depend on the manner in which material is presented (extraneous and germane cognitive loads) or its intrinsic nature (intrinsic cognitive loads).

The main purpose for instructional designers is the design of practice and the organisation and the presentation of information. Instructional interventions can alter extraneous and germane cognitive loads (Van Merrienboer & Sweller, 2005). A poorly designed instruction will be reflected in the effort required by the extraneous cognitive load to process it. On the other hand, when the designed instruction contributes to the generation of schemas of knowledge, it will be reflected in the germane cognitive load. In this sense, appropriate instructional designs decrease extraneous cognitive load but increase the germane cognitive load (Sweller et al., 1998). Although instructional designs can affect extraneous and germane cognitive loads, they cannot alter intrinsic cognitive load because that is intrinsic to the material dealt with and also depends on the interaction between elements of information that a learner must process (interactivity) combined with previous knowledge (Sweller et al., 1998; Paas et al., 2003; Van Merrienboer & Sweller, 2005; Costley & Lange, 2017). However, the intrinsic cognitive load is added directly to extraneous cognitive load when planning instructional designs.

Furthermore, instructional designs should be planned to enhance learning, thus enhancing germane cognitive load. To achieve knowledge, some factors such as the way in which information is presented and the learning activities required should be considered (Van Merrienboer & Sweller, 2005). Based on results that show a positive relationship between auditory, visual and total media and germane cognitive load, Costley and Lange (2017) suggest that instructions should focus on ways of enhancing this cognitive load indicating that video lectures with the diverse forms of media may lead to increase it.

2.1.2 Diverse media and auditory learning

Work has been done that indicates that Diverse media and Auditory instructional designs play an important role in how students perceive lectures and how it affects the learning process (Sweller et al., 1998; R. E. Mayer & Moreno, 2003; Kalyuga & Singh, 2016; Costley & Lange, 2017; R. Mayer, 2017; Boyer, Bevilacqua, Susini, & Hanneton, 2017).

Diverse media involves approaches that promote better understanding and enhance the germane cognitive load. It includes different methods such as presenting the same content multiple times using various forms of media (Paivio, 1991; Schmidt-Weigand & Scheiter, 2011; Khl, Scheiter, Gerjets, & Gemballa, 2011; van Genuchten, van Hooijdonk, Schler, & Scheiter, 2014). Also, it involves the use of animations instead of still images (Khacharem, Zoudji, & Kalyuga, 2015; Dindar, Kabak Yurdakul, & nan Dnmez, 2015; Morrison, Watson, & Morrison, 2015) and the use of videos accompanied by collaborative text (Sloan & Lewis, 2014; Yu, Wang, & Su, 2015; Costley & Lange, 2017; Boyer et al., 2017). On the other hand, auditory learning promotes an optimal germane cognitive load through the process of hearing or speaking, based on approaches such as music and sound added to lectures (Costley & Lange, 2017); and discussions and brainstorming during the learning sessions. Although some research indicates that music and sound added to the class contributes to more learning (Sun & Cheng, 2007), other studies suggest that it may increase distractions affecting the comprehension of the content (R. E. Mayer & Moreno, 2003).

2.2 Mental Workload

2.2.1 Foundations

Longo (2016) suggests that Mental Workload (MWL) is the amount of cognitive work that an individual requires to complete a task over time. This concept is invoked when the complexity of that task needs to be measured. Hart (2006) also suggests that Mental Workload is the result of the interaction between factors related to the person, circumstances and the requirements of that task.

The literature indicates that MWL has applications in ergonomics (Fallahi, Motamedzade, Heidarimoghadam, Soltanian, & Miyake, 2016; Chen, Kang, & Lin, 2016; Doebler, Ryan, Shortall, & Maguire, 2017; J.-Y. Zhang, Liu, Feng, Gao, & Zhang, 2017; Boele-Vos, Commandeur, & Twisk, 2017), human behaviours (Wickens, 2008), computer systems (J. Zhang, Yin, & Wang, 2015; Moustafa, Luz, & Longo, 2017;

Gmyzin, 2017; Caywood, Roberts, Colombe, Greenwald, & Weiland, 2017) and other diverse areas (Cinaz, Arnrich, La Marca, & Tröster, 2013; Wu, Xu, & Lin, 2017; Lassalle et al., 2017; Longo & Leva, 2017).

Xie and Salvendy (2000) suggest that mental overload and underload have a negative effect on performance. Longo (2016) also points out that mental underload refers to the stage when individuals may feel frustrated or annoyed whereas mental overload is related to the stage when an individual is confused, decreasing performance and increasing possible mistakes.

2.2.2 Mental Workload measures

The literature indicates that Mental Workload measures can be classified into three main categories (Young, Brookhuis, Wickens, & Hancock, 2015; Longo, 2016; Moustafa et al., 2017; Gmyzin, 2017). These categories are physiological measures, task performance measures, and subjective measures:

1. **Physiological measures** involve the analysis of physiological indicators and responses of the operator's body obtained from electroencephalogram tests, eye tracking and heart rate measures (Moustafa et al., 2017) that are believed to be correlated to MWL (Longo, 2016).

2. **Task performance measures** relate to the time spent completing a task, the time spent reacting to a second task (indirect quantification of MWL) and the errors made on the primary task (Moustafa et al., 2017; Gmyzin, 2017).

3. **Subjective or self-assessment measures** involve the analysis of a subjective feedback provided by the participant interacting with the task and the system (Moustafa et al., 2017). Longo (2016) points out that an accurate judgement of the task and the MWL experience can only be provided by the person interacting with the task. The Mental Workload, in this case, is usually conducted through surveys or questionnaires related to a post-activity, which relies on two commonly used measures such as the NASA Task Load Index (NASA-TLX) (Hart, 2006) and the Workload profile (WP) (Tsang & Velazquez, 1996).

2.2.3 Subjective measures

NASA-TLX

Subjective measure of MWL that was originally aimed to be used for aviation tasks by pilots within the NASA agency (Moustafa et al., 2017), which then was extended to other fields (Rubio, Díaz, Martín, & Puente, 2004; Rizzo, Dondio, Delany, & Longo, 2016; Rizzo & Longo, 2017). The NASA-TLX is built based on six factors (d_i) and weights (w_i) that reflect the perception of the Mental Workload of a task (Moustafa et al., 2017; Gmyzin, 2017). Such factors are Mental Demand, which is related to the mental skills required to complete the task (remembering, thinking, deciding and other); Physical demand, which involves physical activities (pushing, pulling, and other); Temporal demand, which is related to the pressure felt during the task; Performance, which is measured as self-estimated satisfaction of the perform during the task; Effort, which relates to the amount of effort that the participant applied to achieve the task; and Frustration, which measures how uncomfortable the task was. The weights are calculated using a rating binary choice system related to the combination of the factors in pairs (Longo & Dondio, 2015; Longo, 2017). Based on the factors and the weights, the NASA Task Load Index can be calculated as defined in equation (2.1):

$$NASA - TLX_{MWL} = \left(\sum_{i=1}^{6} di * w_i \right) \frac{1}{15} \tag{2.1}$$

Workload Profile

Wickens (2008) proposed Workload Profile as subjective assessment method under the Multiple Resource Theory (MRT). This measure is based on eight dimensions (d_i) required to perform a task identified by Moustafa et al. (2017). Dimensions of processing such as perceptual or central, response, spatial, verbal, visual, auditory, manual responses and speech responses, which obtained from self-report during the performance of the task (Moustafa et al., 2017; Gmyzin, 2017). Based on the dimensions, the Workload Profile can be calculated as defined in equation (2.2):

$$WP_{MWL} = \sum_{i=1}^{8} di \tag{2.2}$$

For a further understanding of the subjective measures of Mental Workload, the reader is referred to Longo (2012, 2011, 2014, 2015a).

2.3 Natural Language Processing

The literature based on Natural Language Processing indicates that there has been an immense growth in this field during the last ten years due to an increase of large amounts of electronic sources, memory and speed of computers and the use of the internet (Goel, 2017). When a data processing system requires the use of the knowledge of the language, it becomes a language processing system (Jurafsky & Martin, 2014). This is the basis for a set of technologies and theories for the analysis of text (Goel, 2017). There are different mathematical and linguistic approaches used for NLP to solve practical problems and direct real-world applications (Manning, Schütze, et al., 1999). Such approaches or key concepts are subdivided into categories such as syntax, semantics, discourse, and speech:

1. **Syntax** focuses on the rules and techniques that dictate the structure of the sentence in languages (Hurwitz, Kaufman, & Bowles, 2015). It involves subcategories such as lemmatization (Kettunen, Kunttu, & Jrvelin, 2005; Han, Shen, Wang, & Liu, 2012); morphological segmentation (Guinard, 2016); part-of-speech tagging (Lv, Liu, Dong, & Chen, 2016); terminology extraction (Lossio-Ventura, Jonquet, Roche, & Teisseire, 2016); and stemming (Kettunen et al., 2005; Han et al., 2012).

2. **Semantics** is the study of the meaning of words and their combination into meaningful sentences, constructions and utterances (Manning et al., 1999). It involves a number of applications which utilise text theoretically and through implementation. Applications such as lexical semantics (McInnes & Pedersen, 2013; Faruqui et al., 2014); collocations (Manning et al., 1999; Gupta, Nenkova,

& Jurafsky, 2007); named entity recognition (Habib, 2008), (Dai, Lai, Chang, & Tsai, 2015); question answering (Garg & Kumar, 2016); and sentiment analysis (Manke & Shivale, 2015).

3. **Discourse** involves the process of the comprehension and production of naturalistic language as it can provide clues that draw the attention to the aspects of the text that should be focused on and remembered (Crossley, Allen, Kyle, & McNamara, 2014). In NLP, it has a number of applications such as automatic summarisation (Gambhir & Gupta, 2017); conference resolution (Sapena, Padró, & Turmo, 2013); and discourse analysis (Lehmann & Guenthner, 1991).

4. **Speech** has begun to merge with Natural Language Processing as both fields are based on common resources such as raw speech and text corpora, annotated corpora, part-of-speech, among others (Jurafsky & Martin, 2014). In NLP, it involves applications such as speech recognition (Lee & Cho, 2016); speech segmentation (Toro, Sinnett, & Soto-Faraco, 2005), (Panda & Nayak, 2016); and text-to-speech (Shadiev, Huang, & Hwang, 2017).

The definitions provided above, provide a general picture of the applications and approaches of NLP. The aim of this section is to contextualize techniques related to those categories with the research question. In particular, techniques of Natural Language Processing for text preprocessing (tokenizing, stop-words removal, lemmatizing and stemming), similarity measures and weighting scheme for words.

2.3.1 Text preprocessing

1. **Tokenizing** is defined as the process of separating a group of string characters (words, phrases, symbols or other meaningful elements) into tokens after the language of that group of string characters has been identified (Hurwitz et al., 2015).

2. **Stop-words removal** is the process of removing auxiliary words like prepositions, pronouns, conjunctions and interjections that have trivial semantic in-

formation. Therefore, it reduces the total number of words in the corpus and increases the accuracy of semantic algorithms (Furlan, Batanovi, & Nikoli, 2013).

3. **Lemmatizing and Stemming** are techniques used in NLP applications to reduce morphological variations of words mapping them to their base forms (Singh & Gupta, 2016). The difference between both is that stemming obtains the basic word form (the stem), whereas lemmatizing determines the canonica, dictionary or citation form (the lemma) of a word (Paredes-Valverde, Valencia-Garca, Rodrguez-Garca, Colomo-Palacios, & Alor-Hernndez, 2016). The main disadvantage of stemming compared to lemmatizing is that suffix removal algorithms cannot stem the alternate inflection of a word that could have been declared in a different verb tense, or in its plural form, whereas algorithms for lemmatization can find the lemma of the word that correspond to the collection of all word forms that have the same meaning. For example, with a word such as ”defined” (past tense of ”define”), the stemming algorithm would return ”defin”, whereas the lemmatizing algorithm would return ”define”. In this sense, without knowing the real base form, it is not possible to resolve the meaning of the word and the use of those algorithms in applications like word-sense disambiguation it is not feasible (Singh & Gupta, 2016).

2.3.2 Semantic similarity

Determining the relatedness of a word/concept has played an important role in applications for summarising texts, detecting duplicate content and plagiarism, in discourse structures (Harispe, Ranwez, Janaqi, & Montmain, 2017). Harispe et al. (2017) define semantic relatedness as the strength of semantic interactions between two terms without having restrictions on the type of semantic links. Thus, two elements can be highly related despite the fact that they are not similar. In the case that two elements share few of their semantic constitutive properties, semantic similarity, a subset of relatedness, it is used to evaluate the semantic interaction between them based on their taxonomic relationship within an ‘is-a’ hierarchy.

Similarity Measures

McInnes and Pedersen (2013) categorized Semantic similarity measures into two groups: Path-based, which is based on the shortest path information; and Information content, which is based on the shortest path information but incorporating the probability of the concept occurring in a corpus of text.

1. **Path-based** involves a number of approaches such as Distance measure (Rada, Mili, Bicknell, & Blettner, 1989); Reciprocal of length of the shortest path (Caviedes & Cimino, 2004); Distance measure with depth of the Least Common Subsumer (LCS), the WUP similarity method (Wu & Palmer, 1994); Path measure with depth and taxonomy (Leacock & Chodorow, 1998); and other ontology-based approaches (Noy, 2004; Nguyen & Al-Mubaid, 2006; Yang, Yang, & Yuan, 2007; Taieb, Aouicha, & Hamadou, 2014).

2. **Information content** involves a Corpus-based approach that uses the information gained from large corpora to measure the semantic similarity between two concepts (Zhu & Iglesias, 2017); and Taxonomy-based which captures the generality and concreteness of a concept by looking at its incoming (ancestors) and outgoing (descendant) links based on its placement within the hierarchy (McInnes & Pedersen, 2013).

WUP similarity method is a path-based similarity measure proposed by Wu and Palmer (1994). It measures the semantic similarity of two concepts as twice the depth of the most specific concept that is a shared ancestor of the two concepts (LCS), divided by the product of the depths of each individual concept. This semantic similarity measure can be calculated as defined in equation (2.3):

$$sim_{wup} = \frac{2 * depth(lcs_{(c_1,c_2)})}{depth(c_1) + depth(c_2)} \tag{2.3}$$

A large lexical database of English words, WordNet (Miller, 1995), makes use of the Wu and Palmer (1994) method to measure semantic similarity.

WordNet groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms (synsets) (Faruqui et al., 2014). The literature makes reference to WordNet to determine the semantic similarity between words and concepts as it organizes nouns and verbs into hierarchies relations of ”is-a” (Manning et al., 1999; Budanitsky & Hirst, 2006; Taieb et al., 2014; Faruqui et al., 2014; Zhu & Iglesias, 2017). This lexical database can be used for synonym search, for example, as applied in the work of Paredes-Valverde et al. (2016) for querying linked data using NLP.

Synonym search approaches have been proposed based on processing word-level terms (Taboada, Rodriguez, Gudivada, & Martinez, 2017). New synonyms are created from multi-word phrases by replacing one or more words with known synonyms (Allones, Martinez, & Taboada, 2014). A mismatch can happen when a word which is a synonym of another is within a document but is not mentioned in it, thus, reducing the effectiveness of a system and limiting expressiveness by restricting vocabulary (Paredes-Valverde et al., 2016).

2.3.3 Weighting scheme for words

Gupta et al. (2007) suggest that attention has been given to multi-document summarisation in response to a complex user query. Summaries can be of two types: generic or query-focused (Gambhir & Gupta, 2017). Generic summarisation gives an overview of the information in documents, whereas a topic or query can determine what information is appropriate for inclusion in the summary. Regarding aspects of the topic-focused scenario, studies have been conducted in word schemes, such as Word probability and Log-likelihood Ratio (LLR) (Nenkova, 2005; Conroy, Schlesinger, & O’Leary, 2006; Nenkova, Vanderwende, & McKeown, 2006; Gupta et al., 2007).

Relative Frequency ratios

Manning et al. (1999) consider the interpretation of relative frequency ratios as likelihood ratios (an approach to hypothesis testing). Relative frequency ratios between two words can be used to determine the importance of a word that is characteristic of a corpus (core text) when compared to a background corpus (contrastive corpus) (Damerau, 1993). It can be measured as the relative frequency of a word in the core text (Eq. (2.4)), divided by the relative frequency of the word in the contrastive text

(Eq. (2.5)). This measure can be calculated as defined in equation (2.6).

$$FT = \frac{fword_{coreText}}{totalWords_{coreText}} \tag{2.4}$$

$$FC = \frac{fword_{contrasText}}{totalWords_{contrastText}} \tag{2.5}$$

$$RFR = \frac{FT}{FC} \tag{2.6}$$

The purpose of the relative frequency ratio is to compare a general text with a subject-specific text (Manning et al., 1999). In this sense, a relative frequency ratio greater than 1 would indicate that the word is more important to the core text than to the contrastive text. If it is close to 1, then the word is of equal importance to the core text and the contrastive text. Finally, if it is less than one, then the word is more important to the contrastive text than to the core text.

2.4 Related work

Work has been done to measure Mental Workload under instructional designs to enhance health and learning based on applications mainly related to ergonomics (Fallahi et al., 2016; Chen et al., 2016; Doebler et al., 2017; J.-Y. Zhang et al., 2017; Boele-Vos et al., 2017), human factors (Wickens, 2008), machine learning (J. Zhang et al., 2015; Moustafa et al., 2017; Gmyzin, 2017; Caywood et al., 2017) and other diverse areas (Cinaz et al., 2013; Wu et al., 2017; Lassalle et al., 2017). However, collaboration between Mental Workload techniques and Natural language Processing seems to rarely happen. Therefore, to extend the boundaries of the search, related scholarship with particular goals is valuable. For example, Moustafa et al. (2017) applied NASA-TLX and WP to instructional designs to determine the relationship between their indexes and an actual class assigned to the volunteers at the end of the Mental Workload activity. Gmyzin (2017) proposed subjective techniques to compare the performances of these theory-driven measures and of the supervised machine learning models that were trained using the NASA-TLX and WP factors as features to predict a class (indexes of NASA-TLX and WP). Ott et al. (2016) proposed a measure for Mental Workload using multi-modal metrics, namely text, keyboard dynamic data and unstructured linguistics to build a supervised machine learning model. Finally, the studies proposed by McInnes and Pedersen (2013) and by Furlan et al. (2013) related to semantic similarity to text.

2.5 Summary

2.5.1 Gaps in literature

The literature review indicates that the collaboration between Mental workload techniques and Natural language Processing seems to rarely happen. Also, it indicates that Supervised Machine learning has been used to MWL, which reflects an increase application of new technologies to areas usually dominated by theory-driven models such as the subjective measures. Thus, this can be taken as an initial indication that Natural Language Processing can also be used for the same purpose.

It seems that NLP has not yet been applied to MWL under instructional designs and a gap from the studies of Moustafa et al. (2017) and Gmyzin (2017) can provide an idea of what can still be done. In both cases, the hypotheses aimed to determine the relationship between MWL scores (indexes of NASA-TLX and WP) or scores obtained using supervised machine learning models on the one hand and a class applying correlation techniques on the other. The class was collected and designated at the end of the MWL task to each participant or from the NASA-TLX and WP scores. However, from another point of view, it would have been interesting to obtain ratings from text data aiming to measure levels of Mental Workload by applying Natural Language Processing, for example, using keywords collected at the beginning and the end of each task along with the subjective measures.

2.5.2 Research Question

As an appropriate instructional design can increase the germane cognitive load enhancing learning (Sweller et al., 1998), if it is well planned, considering the way in which the information is presented and the learning activities required, it will have a positive impact on individuals' learning, and it might contribute to achieving an optimal Mental workload. Thus, the content of an instrument design translated into text and keywords of activity given under that instrument design can be processed using NLP techniques and then analysed to measure how those keywords are related to the Mental Workload in terms of learning. Accordingly, this research will seek to determine if there is a relationship between Mental Workload and Relative Frequency Ratios of keywords gathered from students during third-level classes under different topics and instructional designs?

Chapter 3

Experiment design and methodology

This chapter provides a definition of the hypotheses necessary to answer the research question. It also involves software selection, data understanding, data preparation, model design, evaluation and hypotheses testing and strengths and limitations of the design approach (see figure 3.1)).

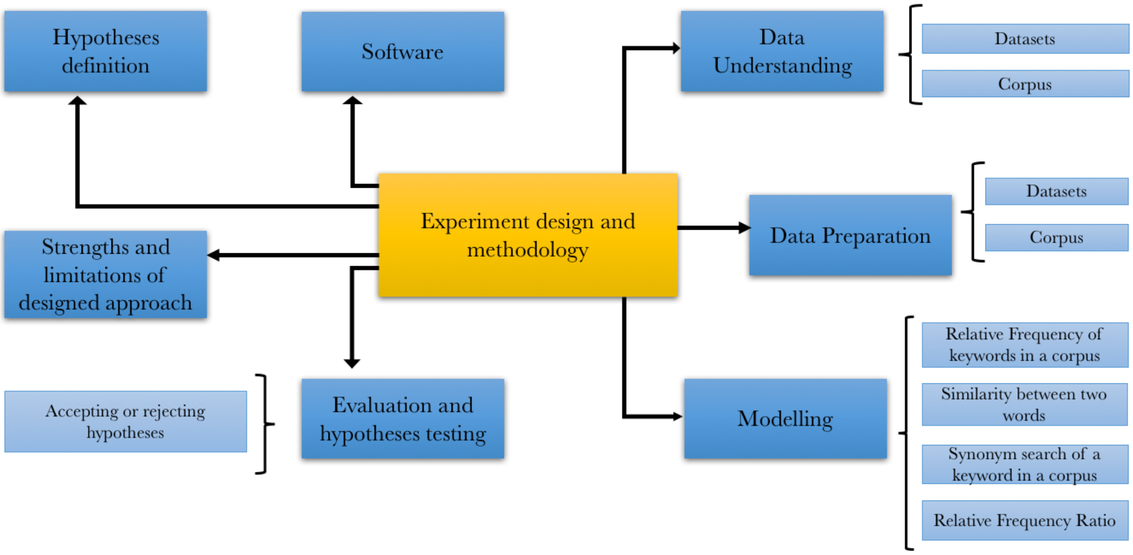

Hypotheses definition
Software
Data Understanding
Datasets
Corpus
Experiment design and methodology
Data Preparation
Datasets
Corpus
Strengths and limitations of designed approach
Relative Frequency of keywords in a corpus
Similarity between two words
Accepting or rejecting hypotheses
Evaluation and hypotheses testing
Modelling
Synonym search of a keyword in a corpus
Relative Frequency Ratio

Figure 3.1: The process of experiment design and methodology which involves an overview of the approaches necessary to answer the research question

The first section begins with the context that permits the formulation of the hypotheses that aims to answer the research question. Then, the hypotheses definition is outlined.

The software section involves the selection criteria of the tools that will be used to conduct each part of the experiments. Thus, the software are presented along with their tasks to be performed.

The data understanding section aims to identify data quality problems and to discover insights into the data. It describes a number of approaches chosen to analyse the NASA-TLX and WP datasets. Also, it presents the analysis and explanation of four core texts, namely: Science, the Scientific Method, Planning Research, and Literature Review. It concludes with the analysis and selection of a contrastive corpus. For the data understanding of the corpus, two pseudo codes are proposed.

The data preparation section presents the steps necessary to solve possible data quality problems such as missing values, outliers, abbreviations, misspellings and assessment of normality during feature generation of the subjective measures of Mental Workload. Also, the steps for feature reduction of the set of keywords features of NASA-TLX and Workload Profile datasets are presented. Finally, during this section five pseudo codes are proposed.

The modelling part aims to determine the importance of a keyword that is characteristic of a core text when compared to a contrastive corpus. It involves the calculation of the similarity between two words, a synonym search and the calculation of Relative Frequency of keywords in a corpus. Finally, the steps to calculate the Relative Frequency Ratios of keywords are presented. During this section, four pseudo codes are proposed.

The evaluation and hypothesis testing section presents the selection of the statistical technique most suitable to address the research question. Then, different possible scenarios are analysed to accept or reject the hypotheses.

Finally, the last section presents the strengths and limitations of the designed approach.

3.1 Hypotheses definition

Between 2015 until 2017, Mental workload experiments were conducted by Dr Luca Longo during third-level classes at Dublin Institute of Technology (DIT). During those experiments four different topics were taught using three instructional designs namely, traditional class, Video-delivery and Video-collaborative. During the sessions, the NASA-TLX and Workload Profile questionnaires were used to collect factors and weights necessary to measure MWL using their mathematical equations (see chapter 2). Along with the questionnaires, keywords were also collected from the participants during the pre-task and post-task. Under those circumstances, the data collected from those sessions permit the formulation of hypotheses that can answer the research question. Based on that, the following two hypotheses are presented (see figures 3.2 and 3.3):

- $H1$: There is a statistically significant relationship between Mental Workload (MWL) and Relative Frequency Ratios (RFR) of keywords gathered from students during third-level classes under different topics.

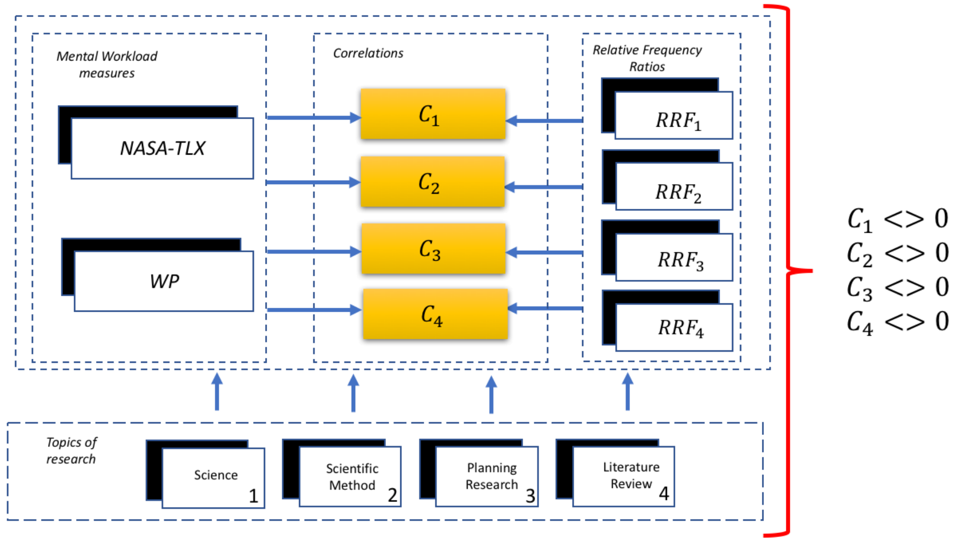


Figure 3.2: Hypothesis definition H1

- $H2$: The strength of the correlation between MWL measures and RFR scores using Video-collaborative (C_3) is greater than the correlation related to the video-delivery approach (C_2) which is also greater than when using a traditional approach (C_1).

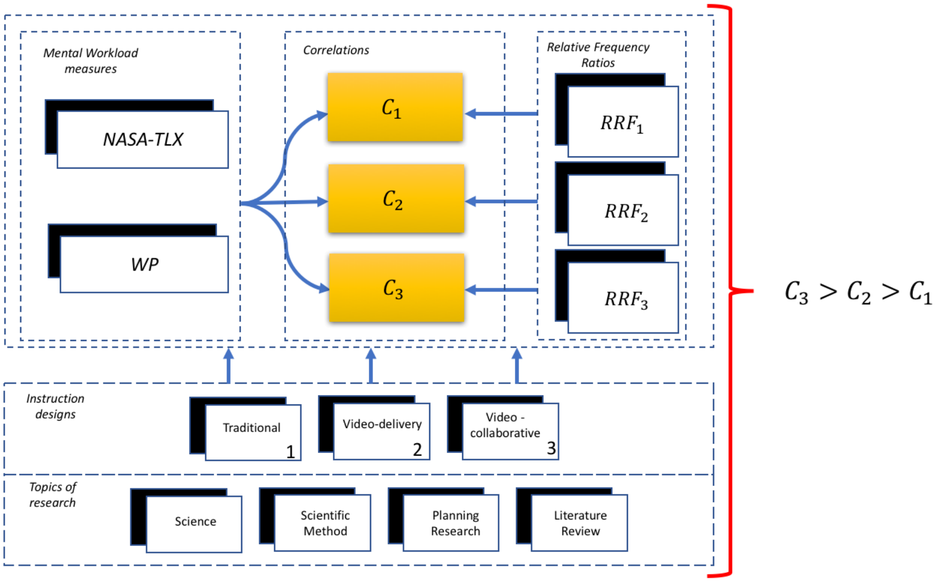


Figure 3.3: Hypothesis definition H2

The content of instructional designs (topics) translated into text along with keywords of activity given under those instrument designs can be analysed to measure how those keywords are related to the Mental Workload. In this sense, the first hypothesis is based on the assumption that Relative Frequency Ratios obtained from the keywords collected during the experiments of MWL at third level sessions and the topics in form of text data provide insights of the MWL activity.

The second hypothesis is an extension of the first assumption but focuses on the instructional designs in terms of the acquisition of learning. Diverse forms of instructional designs approaches promote better understanding and enhance the germane cognitive load. If an instructional design is properly developed and planned, it will have a positive impact on individual's learning and it might contribute to achieving an optimal Mental Workload. During the MWL experiments, topics were given using three instructional designs namely, traditional class, Video-delivery and Video-collaborative. The traditional class involved the presentation of slides along with a lecturer's speech. The video-delivery approach involved the use of videos where the lecturer was pre-recorded explaining the topics along with use of images and text. The video-collaborative approach implemented the video-delivery approach along with group activities to promote discussions supported with the PDF of the slides of the traditional class. In this sense, the hypothesis H2 assumes that the strength of the correlation between MWL measures and RFR scores using Video-collaborative is greater than when using the other two instructional designs, being also the strength of the correlation related to the video-delivery approach greater than when using the traditional approach. The use of Video and collaborative activities as through the process of hearing and seeing and the interaction between the participants can promote an optimal germane cognitive load and it might outperform the video-delivery and the traditional approaches because it is a combination of the former and also it adds discussions between the participants supported with the PDF of the slides of the traditional approach. At the same time, the video-delivery approach might outperform the traditional class because a video of the lecturer supported with images and text, instead of a lecturer's speech supported with slides, might have engaged the participants more.

3.2 Software

Due to a wide range of specific issues that might arise from the data understanding section, this research will conduct each part of the experiments using the most appropriate tool for it based on the criteria that some tools perform differently depending on the activity thus providing different level of efficiency (Gmyzin, 2017).

IBM SPSS statistics has been chosen for the data exploration of the NASA-TLX and Workload Profile datasets and for the statistical analysis of the experiments. Its selection is based on the way it manages data and performs analysis in one go providing a range of techniques to conduct descriptive statistics, test of normality, and statistical tests to determine the relationship between variables.

Python Programming Language has been also selected for data understanding, data preparation, modelling and data visualisation as it provides a range of libraries that facilitate the use of Natural Language Processing techniques. Python will be used for the data analysis of the corpus texts and keyword features. Furthermore, its use is aimed to obtain measures such as the total number of tokens and distinct words in the core texts, percentage of lexical richness, percentage of stop-words and other measures. In data preparation, it will be applied to manage missing values, for feature selection, tokenizing, stop-words removal and lemmatization. For data modelling, it will be used to determine the Relative Frequency Ratios of the keywords features. Finally, for data visualization, Python will be used to present graphs, namely: histograms and box-plots to facilitate the understanding of the results.

3.3 Data Understanding

3.3.1 Datasets

This step will begin with the process of inspecting the NASA-TLX and Workload Profile datasets to identify data quality problems and to discover first insights into the data. The approaches that will be taken to conduct this exploratory analysis are based on the type of features that those datasets have.

NASA-TLX dataset

The table 3.1, NASA-TLX dataset features, indicates that this dataset is divided into five main groups that vary in type. *Instruction Designs*, *Topics* and *Keywords pre and post tasks* are categorical features with categories based on strings. On the other hand, *Dimensions* is a range continuous feature and *Pairwise comparison* is a binary categorical feature. In this sense, an analytic Base Table (ABT) will be built for the *Dimensions* and *Pairwise comparison* features. It will include the number of records, Minimum, Maximum, Mean and Standard Deviation. A table of frequencies will be created for the *Instruction Designs* and *Topics* features that will include the count and percent of frequency measures. Also, a table of frequencies will be built for the *Keywords pre and post tasks* features that will contain the count and unique values.

Workload Profile dataset

The table 3.1, Workload Profile dataset features, indicates that this dataset is divided into four main groups that vary in type. *Instruction Designs*, *Topics* and *Keywords pre and post tasks* are categorical features with categories based on strings while *Dimensions* is a range continuous feature. In this sense, an analytic Base Table (ABT) will be built for the *Dimensions* features. It will include the number of records, Minimum, Maximum, Mean and Standard Deviation. A table of frequencies will be created for the *Instruction Designs* and *Topics* features that will include the count and percentage of frequency measures. Also, a table of frequencies will be built for the *Keywords pre and post tasks* features that will contain the count and unique values.

Table 3.1: NASA-TLX and WP datasets' features ($R = Range, C = Categorical, B = Binary, S = String$)

NASA-TLX features

#	Name	Type
Instruction designs		
1	intru_design	C-S
Topics Taught		
1	topic	C-S
Dimensions (d)		
1	NASA_Mental	R
2	NASA_Physical	R
3	NASA_Temporal	R
4	NASA_Performance	R
5	NASA_Frustration	R
6	NASA_Effort	R
Pairwise comparison (pc)		
1	NASA_TempFrus	C-B
2	NASA_PerMen	C-B
3	NASA_MenPhy	C-B
4	NASA_FrusPer	C-B
5	NASA_TempEffort	C-B
6	NASA_PhyFrus	C-B
7	NASA_PerTemp	C-B
8	NASA_MenEffort	C-B
9	NASA_PhyTemp	C-B
10	NASA_FrustEffort	C-B
11	NASA_PhyPerf	C-B
12	NASA_TempMen	C-B
13	NASA_EffortPhy	C-B
14	NASA_FrustMen	C-B
15	NASA_PerfEffort	C-B

Keywords pre and post tasks

#	Name	#	Name	Type
1	NASA_k1_pre	1	NASA_k1_post	C-S
2	NASA_k2_pre	2	NASA_k2_post	C-S
3	NASA_k3_pre	3	NASA_k3_post	C-S
4	NASA_k4_pre	4	NASA_k4_post	C-S
5	NASA_k5_pre	5	NASA_k5_post	C-S
6	NASA_k6_pre	6	NASA_k6_post	C-S
7	NASA_k7_pre	7	NASA_k7_post	C-S
8	NASA_k8_pre	8	NASA_k8_post	C-S
9	NASA_k9_pre	9	NASA_k9_post	C-S
10	NASA_k10_pre	10	NASA_k10_post	C-S
11	NASA_k11_pre	11	NASA_k11_post	C-S
12	NASA_k12_pre	12	NASA_k12_post	C-S
13	NASA_k13_pre	13	NASA_k13_post	C-S
14	NASA_k14_pre	14	NASA_k14_post	C-S
15	NASA_k15_pre	15	NASA_k15_post	C-S

Workload Profile features

#	Name	Type
Instruction designs		
1	intru_design	C-S
Topics Taught		
1	topic	C-S
Dimensions (d)		
1	WP_solving_deciding	R
2	WP_response_selection	R
3	WP_task_space	R
4	WP_verbal_material	R
5	WP_visual_resources	R
6	WP_auditory_resources	R
7	WP_manual_response	R
8	WP_speech_response	R

Keywords pre and post tasks

#	Name	#	Name	Type
1	WP_k1_pre	1	WP_k1_post	C-S
2	WP_k2_pre	2	WP_k2_post	C-S
3	WP_k3_pre	3	WP_k3_post	C-S
4	WP_k4_pre	4	WP_k4_post	C-S
5	WP_k5_pre	5	WP_k5_post	C-S
6	WP_k6_pre	6	WP_k6_post	C-S
7	WP_k7_pre	7	WP_k7_post	C-S
8	WP_k8_pre	8	WP_k8_post	C-S
9	WP_k9_pre	9	WP_k9_post	C-S
10	WP_k10_pre	10	WP_k10_post	C-S
11	WP_k11_pre	11	WP_k11_post	C-S
12	WP_k12_pre	12	WP_k12_post	C-S
13	WP_k13_pre	13	WP_k13_post	C-S
14	WP_k14_pre	14	WP_k14_post	C-S
15	WP_k15_pre	15	WP_k15_post	C-S

3.3.2 Corpus

Core texts

The Mental workload experiments were conducted by Dr Luca Longo during third-level classes at Dublin Institute of Technology (DIT) under four different topics. Namely, Research Methods and Computer Science, The Scientific Method, Planning Research and Literature review. The lessons were previously translated from Power-point format without using images to a storytelling format. Thus, the third level classes were taught based on the core texts in a storytelling format using three main instructional designs namely, traditional, Video-delivery and Video-collaborative.

1. **Science** is a corpus that has 1888 tokens. The topic involves the definition of Science, the types of Science, its origins, famous scientists, interest of a scientist in research, the relationship between Science and Engineering and the definition of Computer Science.

2. **The Scientific Method** is a corpus that has 2348 tokens. It involves different research methods for Computer Science. It begins with the definition of method and the history of scientific method. Then, it covers approaches for establishing scientific knowledge, its history, famous scientists and their methodologies. Finally, it provides the elements of scientific method, testing hypotheses, problems of inductivism and considerations to take when defining a scientific method.

3. **Planning Research** has a size of 879 tokens. It begins with a concept widely used in research, which formulates the questions what, how, where, who, when and why. Then, it covers approaches for idea generation (Top-down and Bottom-up). Finally, it provides a path of how to start the formulation of a problem.

4. **Literature Review** is a corpus that contains 2353 tokens. It defines a literature, and outlines the contribution of a literature review in defining a specific thesis, problem or research question. Also, it provides tips on writing a literature review for an article, a chapter, or a book. Finally, the literature review and research questions are linked and the formulation of a hypothesis is explained.

Contrastive text

As suggested by Damerau (1993), to determine the importance of a word that is characteristic of a corpus (core text), it is necessary to compare it to a background corpus. In this sense, this research selected an academic topic unrelated to the four core texts explained above and an academic topic with a size (number of tokens) close to the average of the size of the core texts. Based on that, a fragment from the work of Costley and Lange (2017) has been chosen and converted from a PDF format to a '.txt' format. The topic is focused on the effects of instructional designs in Cognitive Load. The fragment of the article includes the introduction, theoretical background, total diversity learning theories, and empirical research supporting total diversity. In size, the corpus has 2044 tokens, thereby following the criteria defined in this mater.

The sizes of the four Core texts and Contrastive text were obtained from Python based on algorithm 1. This analysis was carried out as an initial understanding of the corpus and to be able to select the contrastive text. The algorithm uses the NLTK library for Natural Language Processing where the sequence of characters (tokens), treated as a group, are obtained and then counted.

Algorithm 1 Size of a corpus

```
Import word_tokenize from the nltk library
Read txt file under encoding=”utf-8” given the path
Assign content of the file to an object
Get tokens using word_tokenize applied to the object
Get length of tokens. $Eq(A)$
```

Approaches

The analysis will begin determining how many distinct words the corpus contains to see the number of tokens without duplicates. Then, the lexical richness will be calculated to evaluate the percentage of distinct words and how many times in average each word is used. Also, the frequency of the word tokens will be determined to see their distribution across the texts. Finally, the number of stop-words, punctuation symbols, the total number of effective tokens and their percentage will be calculated. To carry out this step, the algorithm 2 has been proposed.

Algorithm 2 Data Understanding of corpus

```
Get length of tokens. $Eq(A)$
Get distinct words in text. $Eq(B)$
Get lexical richness of the text as $Eq(B) * 100/Eq(A)$.
Get frequency distribution of the tokens.
Get top 25 most common tokens.
Get the number of tokens without stop-words. $Eq(C)$
Get the number of tokens without stop-words and punctuation symbols. $Eq(D)$
Get percent of stop-words as $(Eq(A) - Eq(C)) * 100/Eq(A)$
Get percent of punctuation symbols as $(Eq(C) - Eq(D)) * 100/Eq(A)$
Get the total number of effective tokens. $Eq(D)$
Get percent of the total number of effective tokens as $Eq(D) * 100/Eq(A)$
```

3.4 Data Preparation

3.4.1 Datasets

This step involves solving possible data quality problems such as missing values, outliers, abbreviations, misspellings and assessment of normality during the generation of the subjective measures of Mental Workload and the reduction of the set of keywords features of NASA-TLX and Workload Profile datasets.

NASA-TLX and Workload Profile (WP)

The NASA Task Load Index and the Workload Profile scores will be calculated using their mathematical equations (see equations (2.1) and (2.2) in chapter 2). For the Workload Profile measure, the calculation will be based on the group of *Dimensions* features of the WP dataset (see table 3.1) using algorithm 3. For NASA-TLX, the calculation will be based on the group of *Dimensions* and *Pairwise comparison* features of the NASA-TLX dataset (see table 3.1) using algorithm 4.

Algorithm 3 Workload Profile (WP)

```
 1: Read csv file as Dataframe given the path.
 2: Create an object for each WP dimension feature ($d_1...d_8$ objects).
 3:                                   ▷ See table 3.1 for reference to the lines 2.
 4: Initialize a counter ($i = 0$).
 5: Initialize a list.
 6: for $i$ in range from 0 to the length of $d_1$ do:
 7:     for  values ($id_1, ..., id_8$) in objects ($d_1, ..., d_8$) do:
 8:         Calculate WP as $id_1 + ... + id_8$
 9:         Append WP to the list from line 5.
10:         Increment counter $i = i + 1$
11:     end for
12:     if $i$ equal to the length of $d_1$ is true then
13:         break
14:     end if
15: end for
16: Covert list with WP scores to Dataframe
17: Merge WP Dataframe to the original Dataframe from line 1.
```

Algorithm 4 NASA-TLX

```
 1: Read csv file as Dataframe given the path.
 2: Create an object for each NASA-TLX dimension feature ($d_1 ... d_6$ objects).
 3: Create an object for each NASA-TLX pairwise comparison feature ($pc_1 ... pc_{15}$ objects).
 4:                                ▷ See table 3.1 for reference to the lines 2 and 3.
 5: Initialize a counter ($i = 0$).
 6: Initialize a list.
 7: for $i$ in range from 0 to the length of $d_1$ do:
 8:     Initialize 6 counters ($w_1 ... w_6 = 0$)
 9:     for values ($id_1, ..., id_6, ipc_1, ..., ipc_{15}$) in objects ($d_1, ..., d_6, pc_1, ..., pc_{15}$) do:
10:         if weight achieved from pairwise comparison is true then:
11:             increment counter ($w_i = w_i + 1$)
12:         end if
13:                 ▷ Lines 10 to 12 refer to determining the weight associated to each
        dimension ($w_1 ... w_6$) verifying if it was achieved applying an if statement for each
        pairwise comparison value ($ipc_1, ..., ipc_{15}$).
14:         Calculate NASA-TLX as $(id_1 * w_1 + ... + id_6 * w_6)/15$
15:         Append NASA-TLX to the list from line 5.
16:         Reset counters ($w_1 ... w_6 = 0$)
17:         Increment counter $i = i + 1$
18:     end for
19:     if $i$ equal to the length of $d_1$ is true then
20:         break
21:     end if
22: end for
23: Covert list with NASA-TLX scores to Dataframe
24: Merge NASA-TLX Dataframe to the original Dataframe from line 1.
```

Assessing normality of NASA-TLX and WP

As parametric techniques, such as Pearson correlation, assume normally distributed scores, the frequency distribution of the NASA-TLX and WP will be evaluated to check the violation of this assumption.

The normality will be assessed based on the analysis of the frequency distribution (histogram) of each feature compared to those shown in figure 3.4. A normal distribution will have characteristics of a bell curve indicating that the mean, mode and median are equal. A positive skewness will indicate a clustering of values at the low level of the graph and a negative skewness will indicate a clustering of values at the high end of the graph. To validate the assessment of normality, the statistical test of Kolmogorov-Smirnov will be conducted (Pallant, 2013). A no-significant value (greater than 0.05) will indicate normality. Otherwise, it will suggest that there is a violation of the assumption of normality.

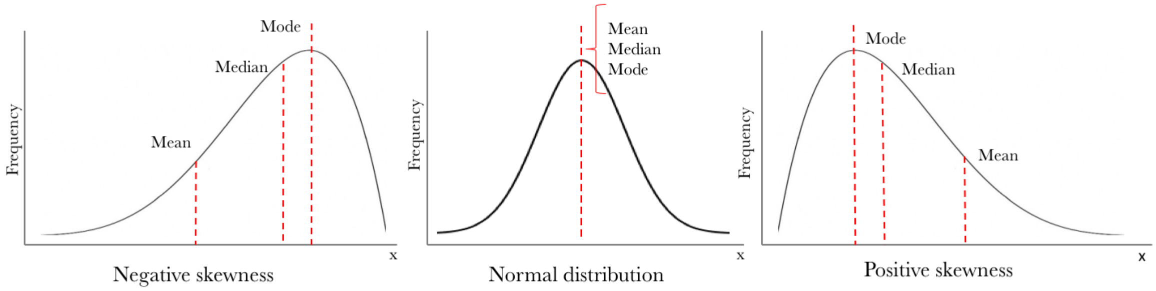


Figure 3.4: Normal Distribution and positive and negative Skewness

Finally, in the case where the scores are positively or negatively skewed, a transformation of the scores will be conducted using mathematical formulas making the variable more 'normally' distributed Pallant (2013).

Keywords

As each dataset has 15 features for pre and post tasks (see table 3.1), those features will be merged to obtain only two main groups (see figure 3.5). To do the aggregation, the missing values must be treated first. During the Mental Workload experiments, participants were asked to write 15 groups of keywords or 15 concepts that described what the participant learned from the teaching session. However, there might be cases where participants left blank spaces as they gave up writing, the task finished because the time was over or the task was completed but the answer was not clear because of the handwriting style. In this sense, the missing values will be imputed with the label 'unknown'. For the reduction of the set of keywords features and the imputation of missing values of the keywords, the algorithm 5 has been proposed.

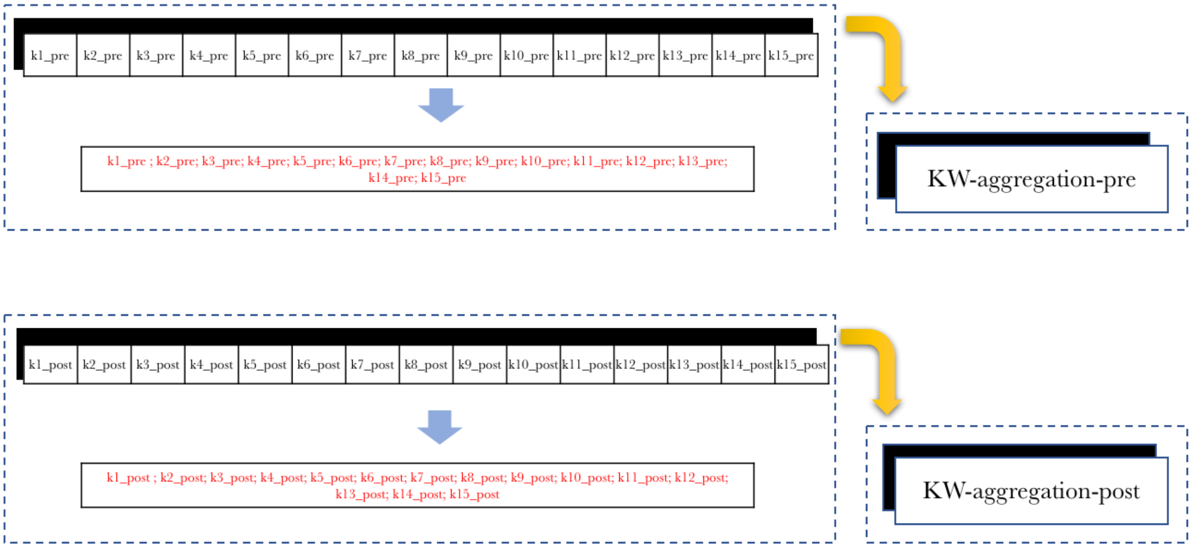

k1_pre
k2_pre
k3_pre
k4_pre
k5_pre
k6_pre
k7_pre
k8_pre
k9_pre
k10_pre
k11_pre
k12_pre
k13_pre
k14_pre
k15_pre
k1_pre ; k2_pre; k3_pre; k4_pre; k5_pre; k6_pre; k7_pre; k8_pre; k9_pre; k10_pre; k11_pre; k12_pre; k13_pre; k14_pre; k15_pre
KW-aggregation-pre
k1_post
k2_post
k3_post
k4_post
k5_post
k6_post
k7_post
k8_post
k9_post
k10_post
k11_post
k12_post
k13_post
k14_post
k15_post
k1_post ; k2_post; k3_post; k4_post; k5_post; k6_post; k7_post; k8_post; k9_post; k10_post; k11_post; k12_post; k13_post; k14_post; k15_post
KW-aggregation-post

Figure 3.5: Reduction of sets of Keywords features to two main features.

Algorithm 5 Aggregation of sets of Keywords features

```
Read csv file as Dataframe given the path.
For the categorical features of Dataframe, fill missing values with 'unknown'.
Create a new feature in the Dataframe named KW-aggregation-pre that will contain the aggregation of the set of 15 features related to the pre task of MWL.
Create a new feature in the Dataframe named KW-aggregation-post that will contain the aggregation of the set of 15 features related to the post task of MWL.
    ▷ The aggregation of the features should be done along axis = 1 separated by ';'.
```

Once *KW-aggregation-pre* and *KW-aggregation-post* are created, the data understanding and data preparation of the new groups of features will be conducted using the algorithm 6.

Algorithm 6 Data understanding and preparation of KW-aggregation features

```
 1: function READ_WORD_GROUP(KW-aggregation as Dataframe):
 2:     Convert KW-aggregation from Dataframe to a list.
 3:     Create an emptied list that will contain calculations for data understanding.
 4:     Create an emptied list that will contain a group of tokens for each participant
    of the MWL task.
 5:     for item in the KW-aggregation list do:
 6:         Get calculations obtained with algorithms 1 and 2.
 7:         Append calculations to a list.
 8:         Convert the list of calculations to a Dataframe where each calculation will
    be a feature.
 9:         Using the Dataframe from line 8 and the list from line 7, create a dictionary
    that will have: total number of tokens, total number of distinct words, average
    of percentage of lexical richness, average of percentage of stop-words, average of
    percentage of punctuation symbols, total number of effective tokens, average of
    percentage of total number of effective tokens.
10:         Convert dictionary to a Dataframe to return calculations
11:         Get object equal to tokens converted to lower-case, without stop-words and
    punctuation symbols, and lemmatized in terms of verbs, adjectives, nouns and
    adverbs.
12:         Append tokens to a list of group of tokens.
13:     end for
14: end function return calculations (line 10) and list of group of tokens (line 12).
```

For the data understanding, the analysis will be conducted in terms of totals and averages for each group of aggregated features. The aim is to calculate the number of distinct words, the average percentage of lexical richness, the average percentage of stop-words, the average percentage of punctuation symbols, the total number of effective tokens and the average percentage of the total number of effective tokens.

The data preparation of the keywords features will begin using the misspelling checker of Microsoft Excel on the datasets to correct any misspelled word. Then, the keywords will be processed to get tokens converted to lower-case, without stop-words and punctuation symbols, and lemmatized in terms of verbs, adjectives, nouns and adverbs (see line 12 of algorithm 6). Also, the features will be inspected in Python using algorithm 7 to find any possible abbreviation that the participants could have used when writing the keywords during the Mental Workload activities. The aim is to find the words in the next and previous position of the possible abbreviations to analyse them to determine whether the possible abbreviation is in fact an abbreviation. Finally, a dictionary of abbreviations will be created with the abbreviations (keys) and their words (values).

Algorithm 7 Inspection of Abbreviations

```
function INSPECT-ABB(keywords):
    for words in keywords do:
        for token in words do:
            if length of token is less or equal to 2 then:
                Find the index of the token in the previous position of token and check if the position is greater or equal to 0 to get the word in the previous position.
                Find the index of the word in the following position of token and check if the position is greater than the length of words −1. If the condition is true, the token is the last element in words. Otherwise, get the word as the following word.
                Print the length of the token, the token, the previous word and the following word.
            end if
        end for
    end for
end function
```

Once the abbreviations have been identified, the algorithm 8 will be used to replace them with the related word. Each abbreviation will be checked comparing the previous and the next words of the possible abbreviation. If an abbreviation is found, it will be replaced with its related word obtained from the dictionary created after the inspection when using algorithm 7.

Algorithm 8 Replace abbreviation for word

```
function REPLACE-ABB(keywordsList,dictionary):
    Get keys of the dictionary and assign them to an object.
    for words in keywords do:
        for token in words do:
            if length of token less of equal to 2 then:
                Find the index of the token in the previous position of token and check if the position is greater or equal to 0 to get the word in the previous position. Otherwise, pass.
                Find the index of the token in the following position of token and check if the position is greater than the length of words −1. If the condition is true, the token is the last element in words. Otherwise, get the following word.
                Print the length of the token, the token, the previous word and the following word.
                Obtain an input from user to verify if the token is an abbreviation (yes 1, otherwise 0). If input is equal to 1, check if the token is within the abbreviations in the dictionary and print the word related to the abbreviation found.
                Obtain an input from user to select the word related to the abbreviation (yes 1, otherwise 0). If input is equal to 1, replace abbreviation for the related word (value in the dictionary). If there are two or more possible words for the abbreviations, ask the user to select one of the options.
            end if
        end for
    end for
    return the list of keywords
end function
```

3.4.2 Corpus

For data preparation, the core and contrastive texts will be processed to get tokens converted to lower-case, without stop-words and punctuation symbols, and lemmatized in terms of verbs, adjectives, nouns and adverbs. The aim is to obtain outputs for each corpus that will be used during the modelling part, namely: a list that contains the tokens of the corpus, another list with the tokens and their individual counts, and the total number of effective tokens. To achieve this task, algorithm 9 has been proposed.

Algorithm 9 Data preparation of a corpus

```
 1: function PROCESS_CORPUS(path):
 2:     Read txt file under encoding=”utf-8” given the path.
 3:     Assign content of the file to an object.
 4:     Get calculations obtained with algorithms 1 and 2.
 5:     Get object equal to tokens converted to lower-case, without stop-words and
        punctuation symbols, and lemmatized in terms of verbs, adjectives, nouns and
        adverbs.
 6:     Use a counter that stores the elements of the object (line 5) as dictionary keys
        and their counts as dictionary values.
 7:     Initialise two lists $list1$ and $list2$
 8:     for element, count in counter (line 6) do:
 9:         Append element to $list1$ and element and count to $list2$.
10:     end for
11: end function return tokens ($list1$), tokens and count ($list2$), total number of
    effective tokens (from line 4).
```

3.5 Modelling

The modelling part aims to determine the importance of a keyword that is characteristic of a core text when compared to a contrastive corpus. It will be measured as the relative frequency of a keyword in the core text (FT), divided by the relative frequency of the keyword in the contrastive text (FC) as defined on the equation (2.6) in chapter 2. As shown in figure 3.6, there are n keywords for each participant of the Mental Workload activity. For each keyword, a Relative Frequency Ratio (RFR) will be calculated. Then, the average of the Relative Frequency Ratios (RFRavg) for each participant will be obtained.

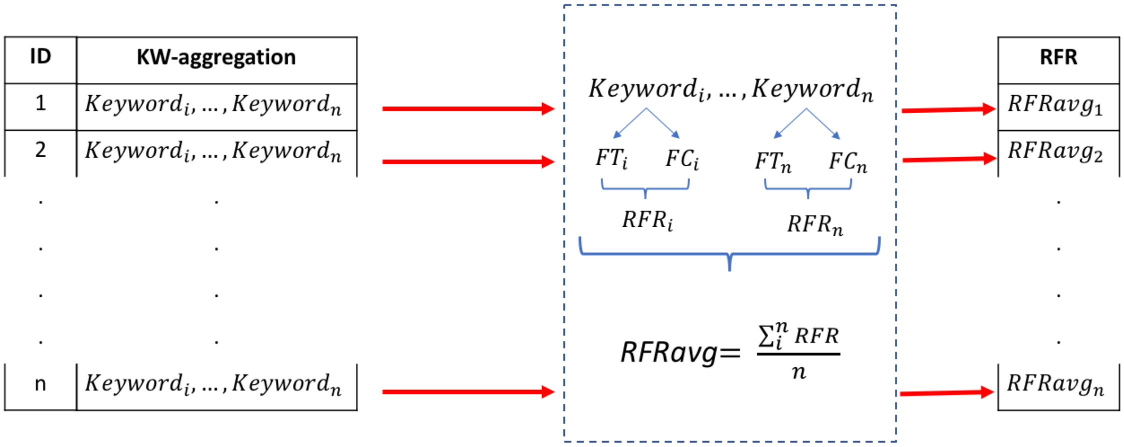

ID
KW-aggregation
1
$Keyword_i, \dots, Keyword_n$
2
$Keyword_i, \dots, Keyword_n$
n
$Keyword_i, \dots, Keyword_n$
$Keyword_i, \dots, Keyword_n$
FT_i
FC_i
FT_n
FC_n
RFR_i
RFR_n
$RFRavg = \frac{\sum_i^n RFR}{n}$
RFR
$RFRavg_1$
$RFRavg_2$
$RFRavg_n$

Figure 3.6: Relative frequency ratio of keywords

Relative frequency of a keyword in a corpus

For the calculation of FT and FC, the frequency of a keyword in the corpus will be divided by the total number of words of the corpus. During this process, a mismatch can happen when a keyword that is a synonym of a word that is within the corpus is not mentioned on it, affecting the effectiveness of the model. To avoid this issue, a synonym search will be conducted based on the similarity of two words under the WUP path-based similarity measure. This task will be carried out using the lexical databased of English words, WordNet, which makes use of Wu and Palmer (1994) method.

1. **Similarity between two words**

   WordNet returns a set of cognitive synonyms (synsets) organized into hierarchies relations. Thus, the semantic similarity of two words will be obtained comparing their sets of synsets applying the WUP measure to each synset. A WUP similarity value will vary from 0 to 1. The minimum value will indicate that the words are neither equal nor synonyms of each other. Otherwise, a similarity of 1 will indicate that the words are equal or have the same meaning. As a word can have multiple synonyms, the maximum of the WUP similarity value will be determined from the synsets found, thus, returning the synonym most similar. To achieved this task, algorithm 10 has been proposed.

2. **Synonym search of a keyword in a corpus**

   To check if a keyword is a synonym of a word that is within the corpus, the keyword has to be compared to each word using the search of the similarity of two words (algorithm 10). In this scenario, each possible synonym and its WUP similarity value will be compared using the maximum to obtain the synonym most similar to the keyword. To achieve this task, algorithm 11 has been proposed.

With the implementation of the synonym search of a keyword in a corpus based on the similarity of two words, it will be possible to determine the relative frequencies FT and FC. In this sense, algorithm 12 has been proposed.

Relative frequency Ratio

Once FT and FC are determined using the algorithm 12. For each keyword, a Relative Frequency Ratio (RFR) will be calculated. Then, the average of the Relative Frequency Ratios (RFRavg) for each participant will be calculated. This task will be achieved using algorithm 13.

Algorithm 10 Similarity between two words

```
1: function FIND-SIMILARITY-TWO-WORDS($word1$,$word2$):
2:     Determine set of synonyms ($syns1$) of $word1$.
3:     Determine set of synonyms ($syns2$) of $word2$.
4:     Initialize a list ($similist$ = []).
5:     for synonym ($s1$) in $syns1$ do:
6:         Initialize a counter ($i = 0$)
7:         for $i$ in range from 0 to the length of $syns2$ do:
8:             for synonym ($s2$) in $syns2$ do:
9:                 Check the WUP similarity between $s1$ and $s2$ greater than a threshold value. Then, append to list in line 4 the synonym found $s2$ and the WUP similarity value. Otherwise, keep checking.
10:                ▷ The threshold value for this experiment (line 9) is equal to 0.8.
11:                Increment counter ($i = i + 1$) and check if the counter ($i$) is equal to the length of $syns2$ to finish the search.
12:            end for
13:        end for
14:    end for
15:    Initialize counter ($maximum = 0$).
16:    for WUP similarity value in list($similist$) do:
17:        Find the maximum checking if the WUP similarity value greater than maximum and assign WUP similarity value as maximum when found.
18:    end for
19:    To return the output of the function, check if the maximum value found is different than 0. Thus, return the WUP similarity value equal to maximum and $word2$ as the synonym found. Otherwise, return the WUP similarity value equal to 0 and $word1$.
20: end function
```

Algorithm 11 Synonym search

```
1: function FIND-SYNONYM-KEYWORD-CORPUS($word1$,$corpus-words$):
2:     Initialize a list ($maxlist = []$)
3:     Initialize an object.
4:     for $word$ in $corpus - words$ do:
5:         Apply the function FIND-SIMILARITY-TWO-WORDS (algorithm 10) to check the semantic similarity between the keyword ($word1$) and a word within the corpus ($word$). This function will return a synonym and its WUP similarity value.
6:         Check if the WUP similarity found (line 5) is different than 0. Then, if the condition is true append the synonym and its WUP similarity value to a list ($maxlist$). Otherwise, keep checking.
7:         Initialize a counter ($maximum = 0$)
8:         for WUP similarity value in list($similist$) do:
9:             Find the maximum checking if the WUP similarity value is greater than maximum. If the condition is true, set WUP similarity value as the maximum and assign its synonym to an object (line 3).
10:        end for
11:    end for
12:    To return the output of the function, check if the maximum value found is different than 0. Thus, return the WUP similarity value equal to maximum and the object (line 9) as the synonym found. Otherwise, return the WUP similarity value equal to 0 and the keyword ($word1$).
13: end function
```

Algorithm 12 Relative frequency of a keyword in a corpus

```
function FREQ-KEYWORD-IN-CORPUS(Input-1,Input-2,Input-3):
    ▷ Input-1 refers to the group of keywords (keywods); Input-2 refers to the group of tokens (words) and their frequencies of a corpus (corpusWordsCount) and Input-3 refers to the group of tokens (words) of the corpus (corpusWords).
    Initialize counter ($i = 0$)
    Initialize a list ($lfreq = []$)
    for $i$ in range from 0 to the lengh of in $keywods$ do:
        Initialize a list ($temp\_list = []$)
        for $item$ in $keywods[i]$ do:
            Check if the keyword ($item$) is in the corpus ($corpusWordsCount$). If the condition is true, append the keyword matched and its frequency to a list ($temp\_list$). Otherwise, apply the FIND-SIMILARITY-TWO-WORDS function (algorithm 11) using the keyword ($item$) and words in the corpus ($corpusWords$) to find a possible synonym and its WUP similarity value.
            Get the synonym and its frequency in the corpus ($corpusWordsCount$) and append those elements to the list ($temp\_list$). If a synonym if not found, append to the list ($temp\_list$) the keyword ($item$) and the frequency of the keyword equal to 0.
        end for
        Append the list $temp\_list$ to the list $lfreq$.
        Determine the frequency of each keyword in the corpus using the list $lfreq$ that contains the keywords and their frequencies along with the total number of effective tokens of the corpus. Assign these calculations to a list that will be the output of the function.
    end for
    return a list containing the frequency of each keyword in the corpus.
end function
```

Algorithm 13 Relative frequency ratio averages

```
function RFR-AVERAGE($FT$,$FC$):
    Initialize a counter ($j = 0$).
    Initialize a list ($fList = []$).
    for $j$ in range from 0 to the length of in $FT$ do:
        Initialize a counter ($rfr\_ = 0$)
        For each keyword determine the Relative Frequency Ratio ($RFR$) dividing FT by FC. Then, increment the counter $rfr\_$ adding each Relative Frequency Ratio ($RFR$).
        Determine the average of the Relative Frequency ratios $RFRavg$ dividing the counter $rfr\_$ by the length of the group of keywords for each participant ($FT[j][1]$).
        Append to a list ($fList$) the average of the Relative Frequency ratios ($RFRavg$).
        Reset counter ($rfr\_ = 0$).
    end for
    return a list ($fList$) containing the average of the Relative Frequency ratios $RFRavg$.
end function
```

3.6 Evaluation and hypothesis testing

The nature of the features that are included in this research will determine which statistical technique is suitable to address the research question. For exploring relationships among the MWL measures (NASA-TLX and WP) and the Relative Frequency Ratio scores (RFR), the Pearson correlation parametric technique and the Spearman Correlation non parametric technique have been chosen. Thus, the Pearson correlation coefficient (r) and the Spearman correlation coefficient (rs) will be calculated representing the correlation between MWL and RFR.

To apply the Pearson correlation, the features must be either interval or ratio measurements and must be approximately normally distributed. A linear relationship between the two variables must exist and the outliers should be kept to a minimum or removed. Finally, a homoscedasticity of the data must exist (Pallant, 2013). The coefficient r can range from -1 to +1 where the sign in front will indicate whether as one variable increases, the other increases too (Positive correlation) or as one variable increases the other decreases (Negative correlation) (Pallant, 2013). If the assumptions of the Pearson correlation are violated, the Spearman correlation will be used to determine the strength and direction of the monotonic relationship between the features, thus, evaluating whether the score of one variable increases, so the score of the other variable increases too or whether the score of one variable increases, the score of the other variable decreases. It works with two variables under the assumption that they must be either ordinal, interval or ratio.

The value of r or rs along with the $pvalue$ will be used to accept or reject the hypotheses. As suggested by Cohen (1988), a value between 0.10 and 0.29 will indicate a small correlation, a value between 0.30 and 0.49 will refer to a medium correlation and between 0.50 and 1 to a large correlation. A correlation of 0 will suggest no relationship at all. To obtain an idea of how much variance MWL and RFR share, the coefficient of determination will be calculated squaring r then multiplied by 100 (Pallant, 2013). This value will provide an indication of how the Relative Frequency Ratios of keywords (RFR) help to explain Mental Workload.

Accepting or rejecting H1

A non-significant value ($p - value > 0.05$) of r or rs will indicate that the hypothesis H1 can be rejected, thus, stating that there is not a statistically significant relationship between MWL and RFR of keywords gathered from students during third-level classes under different topics. On the other hand, a significant value of r or rs will indicate that H1 can be accepted, thus, stating that there is a statistically significant relationship between MWL and RFR of keywords gathered from students during third-level classes under different topics.

Accepting or rejecting H2

The hypothesis H2 states that the strength of the correlation between MWL measures and RFR scores using Video and collaborative (C_3) is greater than the strength of the correlation related to the video-delivery approach (C_2) which is greater than when using a traditional approach (C_1). If the assumption $C_3 > C_2 > C_1$ based on the correlation coefficient (r or rs) is not met, the hypothesis H2 will be rejected. Otherwise, it will be accepted.

3.7 Strengths and limitations of designed approach

3.7.1 Strengths

The first major strength is that **this is novel research that proposes algorithms that are a robust integration of Natural Language Processing techniques to contribute to analyse Mental Workload**, which aligns with the increased use of new technologies applied to theoretical approaches. Algorithms will determine distinct words in a corpus, lexical richness, and the distribution of words across a corpus. They will also calculate NASA-TLX and WP scores, identify and replace abbreviations in texts, and process tokens to lower-case, stop-words and punctuation removal and lemmatization of verbs, adjectives, nouns and adverbs. Also, they will find and replace synonyms, and determine the importance of keywords that are characteristic of a core text when compared to a contrastive corpus (RFR).

Another strength of this design is **the moderately wide range of cases to be evaluated**, as the experiments are based on two datasets, NASA-TLX and WP, where each dataset has four topics and three instructional designs, the combination thereof facilitates more analyses, rather than focusing only on one subjective measure, topic and instructional design.

Finally, **this research is a contribution for future work** because the experiments and models could be expanded to continue the recent need to analyse MWL using Supervised Machine Learning data-driven models. The data generated from this project can be used across different experiments in future works.

3.7.2 Limitations

The first major weakness of this research is the **relatively small size of the data sets**. This makes it difficult to find statistically significant results from the data. Although the NASA-TLX and WP questionnaires were collected from 2015 until 2017, the small size of the datasets is due to the difficulties in conducting Mental Workload activities at third-level sessions because they consume a lot of time which is otherwise spend on teaching the required syllabus, the participants' desirability bias, and the potentially low response rate.

Secondly, a major weakness is **the presence of null values of keywords (missing values)**. In particular, there might be cases where participants gave up writing, the task was finished because the time was over, or the task was completed but the answer was not clear because of the handwriting style. As the main analysis is based on the keywords to get their importance in corpus to determine relative frequency ratios, the missing values pose their own challenges to this research by introducing their imputation based on the assignation of the label 'unknown' which has to be added and handled to a stop-word list during the data preparation section.

Finally, another limitation of the designed approach is related to the **misspelling inspection and correction of the text data**, which although will be performed using a property of Microsoft Excel because of its easy application, it could have been assessed through the use of a designed algorithm. However, the time will be allocated and spent on the identification and replacement of abbreviations, synonyms, word disambiguation in terms of verbs, nouns, adjectives and adverbs and for the calculation of the Mental Workload measures and Relative Frequency Ratios.

Chapter 4

Implementation and results

This chapter presents the results of the performed data understanding, data preparation and modelling of the designed research (see figure 4.1).

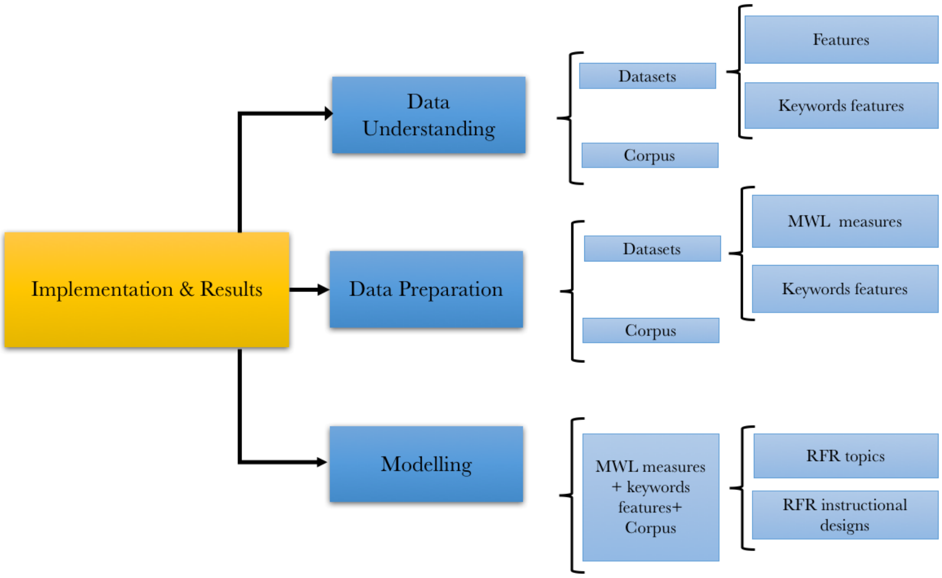


Figure 4.1: Implementation and results process.

The first section begins with the process of inspecting the NASA-TLX and Workload Profile datasets and five corpus, namely: Science, The Scientific method, Planning Research, Literature Review and Contrastive. It is aimed to identify data quality problems and to discover insights into their data, thus, involving a number of approaches chosen in the previous chapter to conduct the analyses.

The data preparation involves the generation of the subjective measures of Mental Workload; the reduction of the set of keywords features of NASA-TLX and Workload Profile datasets; and the data processing of the corpus. Thus, data quality problems namely, missing values, outliers, abbreviations, stop-words and punctuation symbols are handled, along with the assessment of normality of the variables.

Finally, the modelling part aims to determine the importance of a keyword that is characteristic of a core text when compared to a contrastive corpus. For the NASA-TLX and WP datasets under four core texts, three instructional designs and a contrastive corpus, this section begins with the calculation of relative frequencies of keywords. Then, mismatches are avoided using a synonym search, based on the similarity of two words under the WUP path- based similarity measure. Thus, a Relative Frequency Ratio (RFR) is calculated for each keyword and the average of the Relative Frequency Ratios (RFRavg) for each participant is obtained. Finally, the analysis of normality, linearity and homoscedasticity between Mental Workload and Relative Frequency Ratios is conducted.

4.1 Data Understanding

4.1.1 Datasets

This step begins with the process of inspecting the NASA-TLX and Workload Profile datasets, based on the approaches proposed during the experiment design and methodology.

NASA-TLX dataset

An analytic Base Table (ABT) was built for the *Dimensions* and *Pairwise comparison* features (see table 4.1) that includes information for each variable presented as number of records (N), Minimum, Maximum, Mean and Standard Deviation. For the set of *Dimensions*, the features have information from 120 respondents (non missing values), ranging in average from 1.3 (very low) to 18.83 (very high), with a mean of 8.505 and standard deviation of 4.307. On the other hand, the set of *Pairwise comparison*, as binary features, vary from 0 to 1 with the presence of missing values for each feature, meaning that not all participants considered a factor that represented the most important contributor to workload during the teaching sessions. Accordingly, the missing values will be handle during the data preparation as the pairwise comparison features will be used for the calculation of NASA-TLX.

Table 4.1: NASA-TLX: ABT table of Dimensions and partial comparisons features

Features	**N**	**Min**	**Max**	**Mean**	**Sd**
Feature set: Dimensions of NASA-TLX					
NASA_Mental	120	2	20	10.09	3.503
NASA_Physical	120	1	20	6.02	3.949
NASA_Temporal	120	1	20	8.82	3.644
NASA_Performance	120	2	16	8.70	3.499
NASA_Frustration	120	1	17	7.75	4.096
NASA_Effort	120	1	20	9.65	4.148
Feature set: pairwise comparison of NASA-TLX					
NASA_TempFrus	117	0	1	0.23	0.423
NASA_PerMen	117	0	1	0.55	0.500
NASA_MenPhy	116	0	1	0.07	0.254
NASA_FrusPer	117	0	1	0.80	0.399
NASA_TempEffort	115	0	1	0.63	0.486
NASA_PhyFrus	118	0	1	0.50	0.502
NASA_PerTemp	118	0	1	0.47	0.501
NASA_MenEffort	117	0	1	0.32	0.470
NASA_PhyTemp	117	0	1	0.82	0.385
NASA_FrustEffort	117	0	1	0.79	0.406
NASA_PhyPerf	117	0	1	0.91	0.281
NASA_TempMen	117	0	1	0.74	0.443
NASA_EffortPhy	118	0	1	0.09	0.292
NASA_FrustMen	117	0	1	0.81	0.392
NASA_PerfEffort	117	0	1	0.53	0.501

A table of frequencies was created for the *Instruction Designs* and *Topics* features with information presented as the count and percentage of frequency measures (see table 4.2). For the *Instruction Designs* feature, 65 participants (55.2%) attended a traditional class, 46 participants (38.3%) attended a video delivery class, and 9 participants (7.5%) attended a video delivery and collaborative session in the sample, giving a total of 120 participants. On the other hand, for the *Topics* feature, 20 participants (16.7%) attended a class where the topic delivered was *Literature Review*, 31 respondents (25.8%) participated in a class related to *Planning Research*, 36 participants (30%) attended a class related to the *Science* and 33 respondents (27.5%) participated in a class where the topic delivered was *The Scientific Method*, giving a total of 120 participants.

Table 4.2: NASA-TLX: Frequency table of Instruction Designs and Topics features

		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Feature set 4: Instruction designs					
intru_design	traditional	65	54.2	54.2	54.2
	video_delivery	46	38.3	38.3	100.0
	video_and_collaborative	9	7.5	7.5	61.7
	Total	120	100.0	100.0	
Feature set 5: Topic					
topic	literature_review	20	16.7	16.7	16.7
	planning_research	31	25.8	25.8	42.5
	science	36	30.0	30.0	72.5
	the_scientific_method	33	27.5	27.5	100.0
	Total	120	100.0	100.0	

A table of frequencies was built for the *Keywords pre-task and post-task* features with information for each variable presented as the count (see table 4.3). It can be said that there exists the presence of missing values for all keywords features except for *NASA_k1_pre* and *NASA_k2_pre*. There is also a pattern shown in table 4.3 that indicates that as the number of keywords features increase, the number of missing values increase too. That might have happened because participants left blank spaces as they gave up writing or the task finished because the time was over. Accordingly, the missing values will be imputed with the label 'unknown' during the preparation part as proposed in the previous chapter.

Table 4.3: NASA-TLX: Frequency table of keywords pre-task and post-task

	count		**count**
Feature set: keywords pre-task and post-task			
NASA_k1_pre	120	NASA_k1_post	112
NASA_k2_pre	120	NASA_k2_post	112
NASA_k3_pre	116	NASA_k3_post	111
NASA_k4_pre	115	NASA_k4_post	111
NASA_k5_pre	114	NASA_k5_post	109
NASA_k6_pre	112	NASA_k6_post	109
NASA_k7_pre	111	NASA_k7_post	108
NASA_k8_pre	109	NASA_k8_post	105
NASA_k9_pre	107	NASA_k9_post	103
NASA_k10_pre	107	NASA_k10_post	103
NASA_k11_pre	104	NASA_k11_post	101
NASA_k12_pre	100	NASA_k12_post	98
NASA_k13_pre	100	NASA_k13_post	94
NASA_k14_pre	96	NASA_k14_post	90
NASA_k15_pre	94	NASA_k15_post	87

Workload Profile dataset

An analytic Base Table (ABT) was built for the *Dimensions* features (see table 4.4) that includes information for each variable presented as number of records (N), Minimum, Maximum, Mean and Standard Deviation. For the set of *Dimensions*, the features have information from 105 respondents (non missing values), ranging in average from 2.5 to 20, with a mean of 11.224 and standard deviation of 4.003.

Table 4.4: WP: ABT table of Dimensions features

Features	**N**	**Min**	**Max**	**Mean**	**Sd**
Feature set: Dimensions of Workload Profile					
WP_solving_deciding	105	4	20	11.586	3.7497
WP_response_selection	105	1	20	10.71	3.7881
WP_task_space	105	1	20	9.052	4.3434
WP_verbal_material	105	5	20	12.271	3.6971
WP_visual_resources	105	3	20	12.79	3.902
WP_auditory_resources	105	4	20	13.005	3.7879
WP_manual_response	105	1	20	10.37	4.486
WP_speech_response	105	1	20	10.01	4.507

A table of frequencies was created for the *Instruction Designs* and *Topics* features with information presented as the count and percent of frequency measures(see table 4.5). For the *Instruction Designs* feature, 68 participants (64.8%) attended a traditional class, 28 participants (26.7%) attended a video delivery class, and 9 participants (8.6%) attended a video delivery and collaborative session in the sample, giving a total of 105 participants. On the other hand, for the *Topics* feature, 20 participants (19%) attended a class where the topic delivered was *Literature Review*, 28 respondents (26.7%) participated in a class related to *Planning Research*, 28 participants (26.7%) attended a class related to the *Science* and 29 respondents (27.6%) participated in a class where the topic delivered was *The Scientific Method*, giving a total of 105 participants.

Table 4.5: WP: Frequency table of Instruction Designs and Topics features

		Frequency	**Percent**	**Valid Percent**	**Cumulative Percent**
Feature set 4: Instruction designs					
intru_design	traditional	68	64.8	64.8	64.8
	video_delivery	28	26.7	26.7	100
	video_and_collaborative	9	8.6	8.6	73.3
	Total	105	100	100	
Feature set 5: Topic					
topic	literature_review	20	19	19	19
	planning_research	28	26.7	26.7	45.7
	science	28	26.7	26.7	72.4
	the_scientific_method	29	27.6	27.6	100
	Total	105	100	100	

A table of frequencies was built for the *Keywords pre-task and post-task* features with information for each variable presented as the count (see table 4.6). It can be said that there exists the presence of missing values for all keywords features except for *WP_k1_pre*, *WP_k2_pre* and *WP_k3_pre*. There is also a pattern shown in table 4.6 that indicates that as the number of keywords features increase, the number of missing values increases too, similar to the keywords features of the NASA-TLX dataset.

Table 4.6: WP: Frequency table of keywords pre-task and post-task

	count		count
Feature set: keywords pre-task and post-task			
WP_k1_pre	105	WP_k1_post	104
WP_k2_pre	105	WP_k2_post	104
WP_k3_pre	105	WP_k3_post	104
WP_k4_pre	103	WP_k4_post	104
WP_k5_pre	102	WP_k5_post	101
WP_k6_pre	100	WP_k6_post	102
WP_k7_pre	98	WP_k7_post	102
WP_k8_pre	98	WP_k8_post	102
WP_k9_pre	96	WP_k9_post	98
WP_k10_pre	92	WP_k10_post	99
WP_k11_pre	90	WP_k11_post	94
WP_k12_pre	86	WP_k12_post	93
WP_k13_pre	81	WP_k13_post	87
WP_k14_pre	81	WP_k14_post	85
WP_k15_pre	76	WP_k15_post	83

4.1.2 Corpus

The calculations of the descriptive measures of the corpus were obtained using algorithm 1 and 2 proposed in chapter 3. As shown in table 4.7 the information for each corpus is presented as the number of distinct words, lexical richness, frequency of the word tokens, the number of stop-words, punctuation symbols, the total number of effective tokens and the percentage of total number of effective tokens.

Table 4.7: Descriptive of core texts

Measure	**Science**	**The Scientific Method**	**Planning Research**	**Literature Review**	**Contrastive**
Total tokens	1888	2348	879	2353	2044
Total distinct words	745	784	346	686	568
Percent Lexical richness	39%	33%	39%	29%	28%
Percent Stop-words	28.44%	31.77%	37.77%	36.04%	26.71%
Percent other symbols	20.87%	18.40%	17.86%	17.85%	21.53%
Total effective tokens	957	1170	390	1085	1058
Percent effective tokens	50.69%	49.83%	44.37%	46.11%	51.76%

1. **Science**

   The Science corpus has 1888 tokens (words and punctuation symbols) and 745 distinct words (word types and punctuation symbols) which represent 39% (lexical richness) of the total number of words indicating that each word is used two times on average. From the total number of tokens, 28.44% are stop-words and 20.87% represent other symbols, which means that when removing them during the data preparation, the total number of effective tokens will be 50.69% (957 tokens).

Figure 4.2 shows the 25 most frequently occurring word types and punctuation symbols in the Science corpus. The words ‘Science’, ‘Computer’, ’engineering’ and ’Died’, are the most informative in this text. This makes sense because this topic involves the definition of Science, its types, its origins, famous scientists, the relationship between Science and Engineering, and the definition of Computer Science. It can be noticed from figure 4.2 that the rest of the words do not provide enough information about the text as most of them are stop-words and punctuation symbols. Also, there are capitalised words in the text and cases of different word forms that have the same meaning (’computer’ and ‘computers’). Stop-words and punctuation symbols removal, lower-case transformation and lemmatization will be conducted during the data preparation.

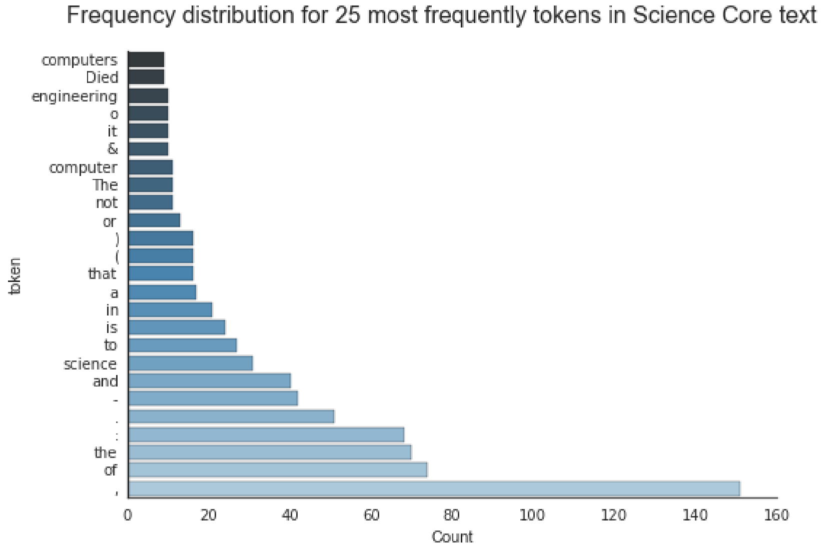


Figure 4.2: Frequency distribution of Science Core text indicating how the top 25 number of word tokens in the text are distributed across the vocabulary items.

2. **The Scientific Method**

The Scientific Method corpus has 2348 words and punctuation symbols and 784 distinct words which represent 33% (lexical richness) of the total number of words indicating that each word is used three times on average. From the total number of tokens, 31.77% are stop-words and 18.40% represent other symbols, which means that when removing them during the data preparation, the total number of effective tokens will be 49.83% (1170 tokens).

Figure 4.3 shows the 25 most frequently occurring word types and punctuation symbols in the Scientific Method core text. The words 'scientific', 'method', 'science' and 'hypothesis' are the most informative in this topic and they occur around 30 times. This makes sense because this topic involves the definition of method, the history of scientific method and its elements, and testing hypotheses techniques. On the other hand, the rest of the words do not provide enough information about the text because most of them are stop-words and punctuation symbols that will be removed in the next section.

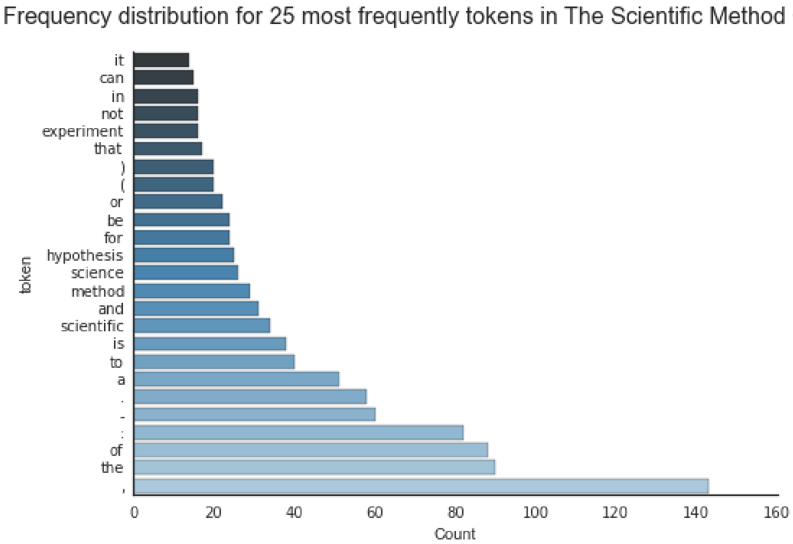


Figure 4.3: Frequency distribution of The Scientific Method Core text indicating how the top 25 word tokens in the text are distributed across the vocabulary items.

3. **Planning Research**

The Planning Research corpus has 879 words and punctuation symbols and 346 distinct words which represent 39% (lexical richness) of the total number of words indicating that each word is used 2 times on average. From the total number of tokens, 37.77% are stop-words and 17.86% represent other symbols, which means that when removing them during the data preparation, the total number of effective tokens will be 44.37% (390 tokens).

Figure 4.4 shows the 25 most frequently occurring word types and punctuation symbols in the Planning Research core text. The words 'Problem' and 'Research' are the most informative words in this topic and they occur around 10 times. This makes sense as the Planning Research topic covers a concept widely used in research and the formulation of a problem. On the other hand, the rest of the words do not provide enough information about the text as most of them are stop-words and punctuation symbols that will be removed in the next section.

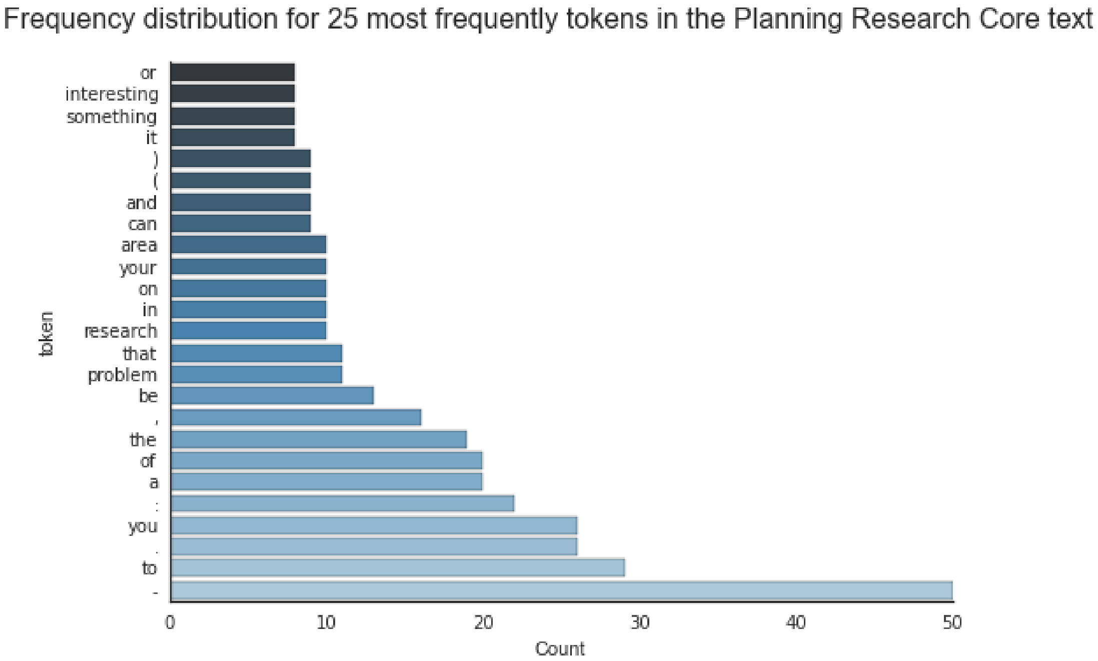


Figure 4.4: Frequency distribution of The Planning Research Core text indicating how the top 25 word tokens in the text are distributed across the vocabulary items.

4. **Literature Review**

The Literature Review corpus has 2353 tokens and 686 distinct words which represent 29% (lexical richness) of the total number of words indicating that each word is used two times on average. From the total number of tokens, 36.04% are stop-words and 17.85% represent other symbols, which means that when removing them during the data preparation, the total number of effective tokens will be 46.11% (1085 tokens).

Figure 4.5 shows the 25 most frequently occurring word types and punctuation symbols in the Literature Review core text. The word 'Research' is the most informative word in this topic and it occurs around 60 times. Also, the words 'review', 'question', and 'literature' provide information about the corpus and they appear around 30 times. This makes sense because this topic involves the definition of what a literature review is and its contribution to define a research question. On the other hand, the rest of the words do not provide enough information about the text as most of them are stop-words and punctuation symbols that will be removed in the next section.

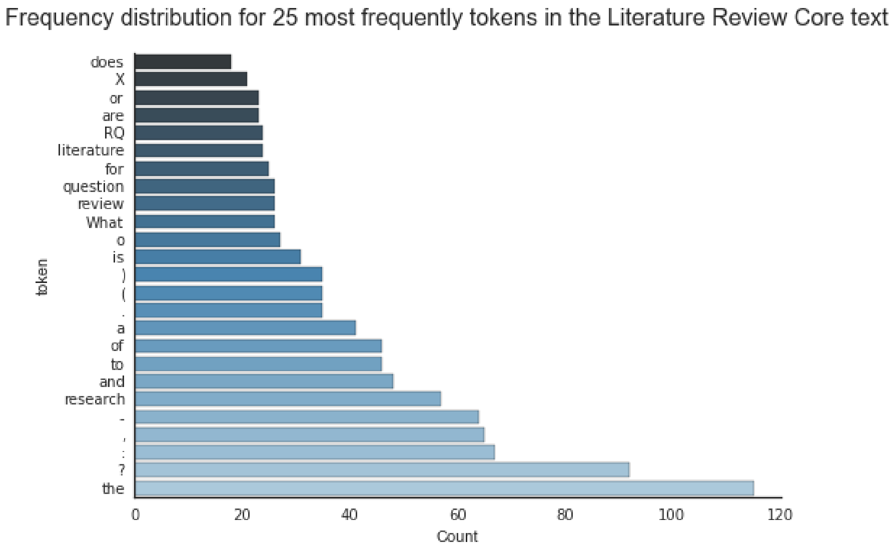


Figure 4.5: Frequency distribution of The Literature Review Core text indicating how the top 25 number of word tokens in the text are distributed across the vocabulary items.

5. **Contrastive**

The Contrastive corpus has 2044 tokens and 568 distinct words which represent 28% (lexical richness) of the total number of words indicating that each word is used 3 times on average. From the total number of tokens, 26.71% are stop-words and 21.53% represent other symbols, which means that when removing them during the data preparation, the total number of effective tokens will be 51.76% (1085 tokens).

Figure 4.5 shows the 25 most frequently occurring word types and punctuation symbols in the Literature Review core text. The words ‘media’, ‘load’, ‘content‘, ‘students’, ‘cognitive’ and ‘learning’ are the most informative words in this topic and they occur around 25 times. This makes sense because the contrastive topic is focused on the effect of instructional designs in Cognitive load on students, including theoretical background and total diversity learning theories. On the other hand, the rest of the words do not provide enough information of the text as most of them are stop-words and punctuation symbols that will be removed in the next section.

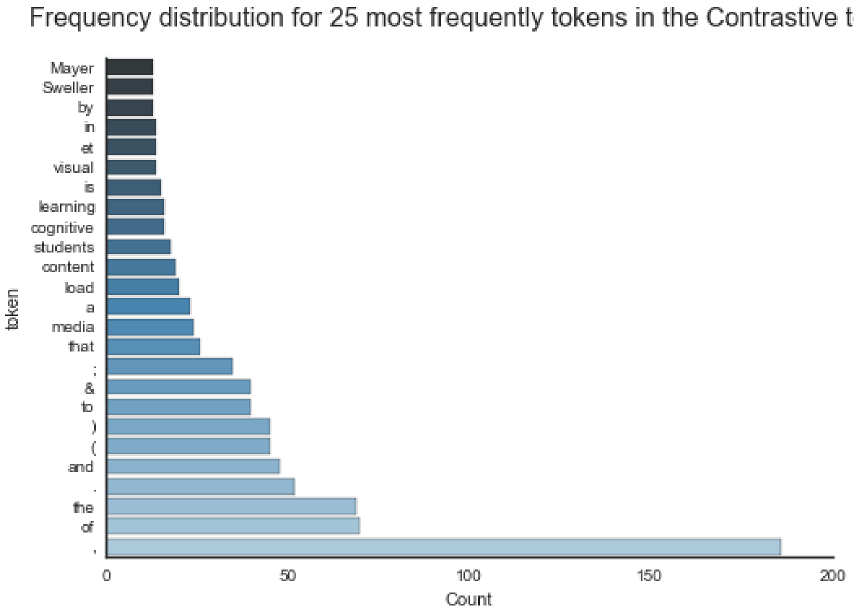


Figure 4.6: Frequency distribution of The Contrastive corpus indicating how the top 25 word tokens in the text are distributed across the vocabulary items.

4.2 Data Preparation

4.2.1 Datasets

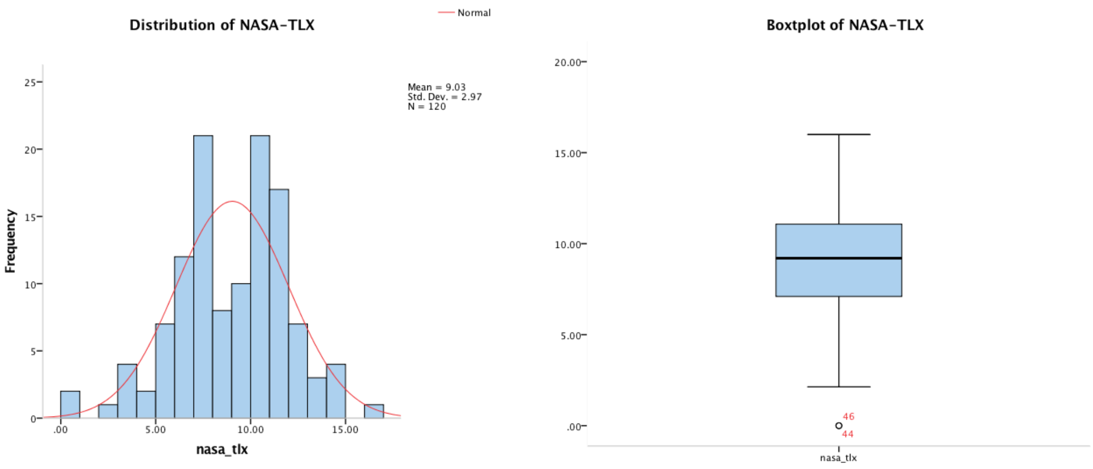
This step involves the feature generation of the subjective measures of Mental Workload and the feature reduction of the set of keywords features of NASA-TLX and Workload Profile datasets. Thus, data quality problems namely, missing values, outliers and abbreviations are handled along with the assessment of normality of the variables.

NASA-TLX

During the data understanding of the NASA-TLX dataset, the set of *Pairwise comparison* features were found with missing values. This means that not all participants considered a factor that represented the most important contributor to the workload during the teaching sessions. For the calculation of NASA-TLX those values would determine the weights that multiply their related dimension. Accordingly, it was decided that the missing values would represent weights equal to 0 because replacing them with the mean value for the variable or excluding them could severely distort the results of the analysis.

For NASA-TLX, the calculation was based on the group of *Dimensions* and *Pairwise comparison* features of the NASA-TLX dataset (see tables 3.1 and **??**) using the algorithm 4 (see chapter 3).

As shown in figure 4.7a, NASAT-TLX has a mean of 9.03 and standard deviation of 2.97 for 120 participants, with an actual shape of distribution that tends to be symmetrical with two peaks in the centre. Looking at the tails of the distribution, there are data points sitting on their own out on the extremes, indicating the presence of potential outliers.



(a) Histogram.

(b) Boxplot.

Figure 4.7: NASA-TLX: Assessment of normality.

The boxplot on the right (figure 4.7b), confirms the analysis of the distribution, indicating the presence of two outliers that extend more than 1.5 box-lengths from the edge of the box (participants with ID equal to 44 and 46). From checking the outliers' scores in the dataset, it was found that the participants (44 and 46) did not assign weights for the dimensions of NASA-TLX. As the weights multiply the dimensions, this produced a measure of Mental Workload equal to 0 (see equation 2.1 in chapter 2). As the outliers turn out to be genuine scores, it was decided to remove them from the data file.

Based on the significant result of the test of normality of Kolmogorov-Smirnov shown in table 4.8 ($0.012 < 0.05$), it can be said that the NASA-TLX does not have a normal distribution.

Table 4.8: NASA-TLX : Test of normality.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
NASA-TLX	0.094	120	0.012

When removing the outliers from the dataset (see figure 4.8), a second test of normality was conducted for 118 participants (see table 4.9). The output of the test indicates an improvement of the distribution of NASA-TLX but a significant result of 0.021 < 0.05 was still obtained.

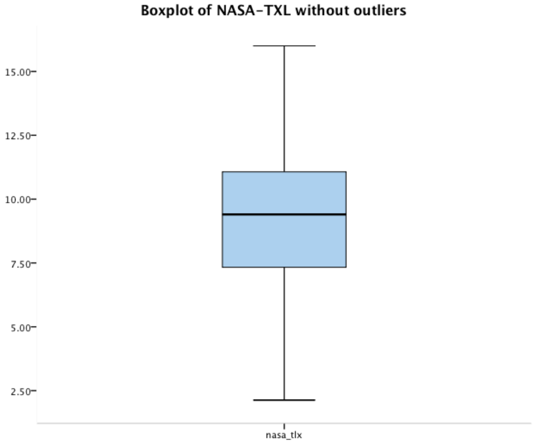


Figure 4.8: NASA-TLX: boxplot without outliers.

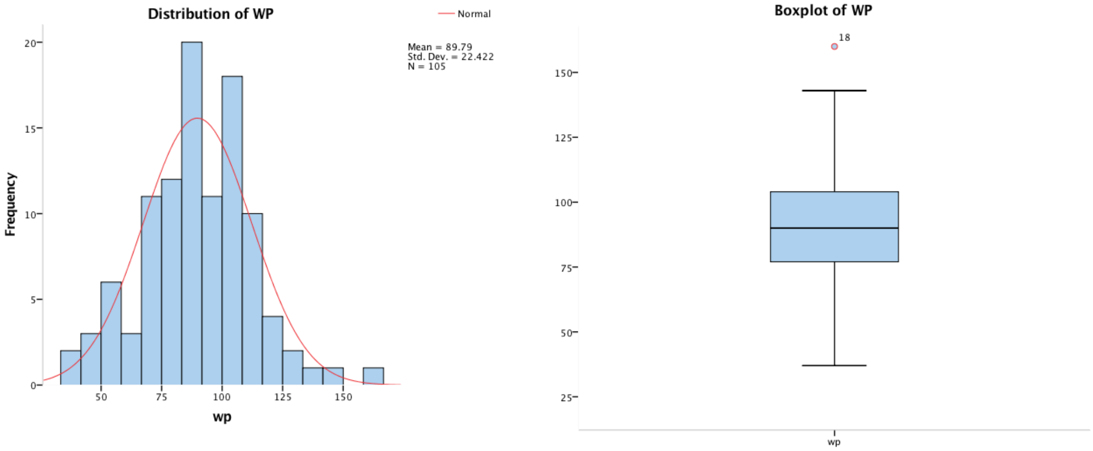
Table 4.9: NASA-TLX : Test of normality (2nd test).

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
NASA-TLX	0.094	118	0.021

WP

For the Workload Profile measure, the calculation was based on the group of *Dimensions* features of the WP dataset (see table 3.1) using the algorithm 3 (see chapter 3).

As shown in figure 4.9a, WP has a mean of 89.79 and standard deviation of 22.42 for 105 participants, with an actual shape of distribution that tends to be symmetrical with peaks in the centre. On examination of the tails of the distribution, there is a data point sitting on its own, out on the right extreme, indicating the presence of a potential outlier.



(a) Histogram.

(b) Boxplot.

Figure 4.9: WP: Assessment of normality.

The boxplot on the right (figure 4.9b), confirms the analysis of the distribution indicating thus the present of one outlier that extends more than 1.5 box-lengths from the edge of the box (participant with ID equal to 18). From checking the outlier's scores in the dataset, it was found that the participant assigned the maximum values for the dimensions of WP. As the dimensions are added to calculate the measure of the Mental Workload activity, it produced the highest value (160) (see equation 2.2 in chapter 2). As the outlier turns out to be a genuine score outside two standard deviations of the mean, it was decided to remove it from the data file.

Based on the non-significant result of the test of normality of Kolmogorov-Smirnov shown in table 4.10 ($0.2 > 0.05$), it can be said that WP has a normal distribution.

Table 4.10: WP: Test of normality.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
WP	0.067	105	0.2

When removing the outlier from the dataset (see figure 4.10), a second test of normality was conducted for 104 participants (see table 4.11). The output of the test indicates that after removing the outlier WP is still normally distributed.

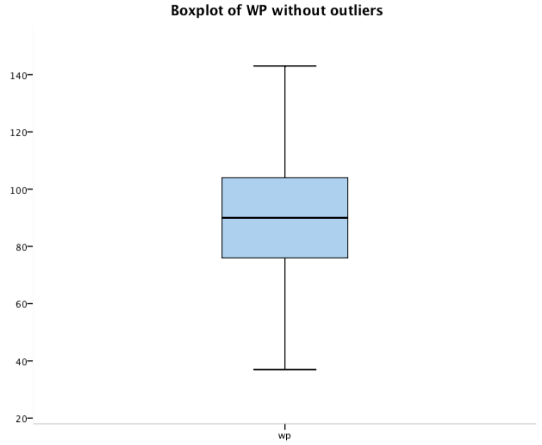


Figure 4.10: WP: boxplot without outliers.

Table 4.11: WP: Test of normality (2nd test).

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
WP	0.070	104	0.02

Keywords

During the data understanding of the NASA-TLX and WP datasets, the set of *keywords pre-task and post-task* features were found with missing values, meaning that some participants left blank spaces as they gave up writing or the task finished because the time was over. As proposed in chapter 3, the missing values were imputed with the label 'unknown'. Then, the features were merged to obtain two main groups for each dataset (see table 4.12). The reduction of the set of keywords features and the imputation of their missing values were conducted using the algorithm 5 (see chapter 3).

Table 4.12: Keywords: Frequency table of keywords pre-task and post-task for NASA-TLX and WP datasets

	count
Features set: Keywords	
NASA_KW_aggregation_pre	118
NASA_KW_aggregation_post	118
WP_KW_aggregation_pre	104
WP_KW_aggregation_post	104

Once the new features were created, the data understanding and data preparation were conducted using the algorithm 6 proposed in chapter 3. As shown in table 4.13 the information for each keyword feature is presented as the number of distinct words, average of the percentage of lexical richness, average of the total number of stop-words, average of the total of punctuation symbols, the total number of effective tokens and the average of the percentage of total number of effective tokens.

Table 4.13: Descriptive statistics of group of sets of Keywords related to the WP pre task.

	NASA_KW_aggregation		**WP_KW_aggregation**	
Measure	**pre**	**post**	**pre**	**post**
Total tokens	5716	5657	5086	5349
Total distinct words	2889	2784	2583	2696
Avg_Percent Lexical richness	50.37%	48.38%	50.65%	50.44%
Avg_Percent Stop-words	9.34%	9.58%	10.73%	12.72%
Avg_Percent other symbols	30.44%	31.49%	30.47%	29.21%
Total effective tokens	3444	3342	2979	3098
Avg_Percent effective tokens	60.22%	58.93%	58.80%	58.07%

For the **NASA-TLX** dataset, the feature related to the keyword pre-task has 5716 tokens and 2889 distinct words which represent on average 50.37% of the total number of keywords. From the total number of tokens on average, 9.34% are stop-words and

30.44% represent other symbols, which means that when removing them during the data preparation, the total number of effective tokens will be 60.22% (3444 keyword tokens). On the other hand, the feature related to the keyword post-task has 5657 tokens and 2784 distinct words which represent on average 48.38% of the total number of keywords. From the total number of tokens on average, 9.58% are stop-words and 31.49% represent other symbols, which means that when removing them during the data preparation, the total number of effective tokens will be 58.93% (3342 keyword tokens).

For the **WP** dataset, the feature related to the keyword pre-task has 5086 tokens and 2583 distinct words which represent on average 50.65% of the total number of keywords. From the total number of tokens on average, 10.73% are stop-words and 30.47% represent other symbols, which means that when removing them during the data preparation, the total number of effective tokens will be 58.80% (2979 keyword tokens). On the other hand, the feature related to the keyword post-task has 5349 tokens and 2696 distinct words which represent on average 50.44% of the total number of keywords. From the total number of tokens on average, 12.72% are stop-words and 29.21% represent other symbols, which means that when removing them during the data preparation, the total number of effective tokens will be 58.07% (3098 keyword tokens).

The data preparation of the keyword features began using the misspelling checker of Microsoft Excel on the datasets to correct any misspelled word. Then, in Python, the keywords were converted to tokens where each token was transformed to lower-case. The label ‘unknown’, assigned to the missing values of the keyword features, was added to the list of stop-words used for the experiments from the NLTK library in Python. The features were inspected using algorithm 7 (see chapter 3) to find any possible abbreviation that the participants could have used when writing the keywords during the Mental Workload activities. The abbreviations found were ‘c’ for ‘computer’,‘s’ for ‘science’ or ‘scientific’,‘cs’ for ‘computer-science’ and ‘vs’ for ‘versus’. From the inspection, a dictionary, with the abbreviations (keys) and their words (values) was created. Then, the abbreviations were replaced with the related word using algorithm

8 proposed in chapter 3. The stop-words and the punctuation symbols were then removed. As a final step for data preparation, the tokens were lemmatized in terms of verbs, adjectives, nouns and adverbs.

Finally, after the data preparation, for the NASA-TLX dataset, the feature related to the keyword pre-task resulted with 3444 effective tokens (60.22%) and the post-task feature resulted in 3342 effective tokens (58.93%). On the other hand, for the WP dataset, the feature related to the keyword pre-task resulted in 2079 effective tokens (58.80%) and the post-task feature resulted in 3098 effective tokens (58.07%).

4.2.2 Corpus

The corpus were processed to get tokens converted to lower-case, without stop-words and punctuation symbols, and lemmatized in terms of verbs, adjectives, nouns and adverbs, using algorithm 9 (see chapter 3). Thus, the Science corpus was left with 957 effective tokens, which represents a 50.69% of its total number of tokens. The Scientific Method corpus resulted with 1170 effective tokens, a 49.83% from its total number of tokens. The Planning Research corpus was left with 390 effective tokens, which represents 44.37% from its total number of tokens. The Literature review corpus resulted in 1085 effective tokens, a 46.11% from its total number of tokens. Finally, the Contrastive corpus was left with 1085 effective tokens, which represents 51.76% of its total number of tokens.

4.3 Modelling

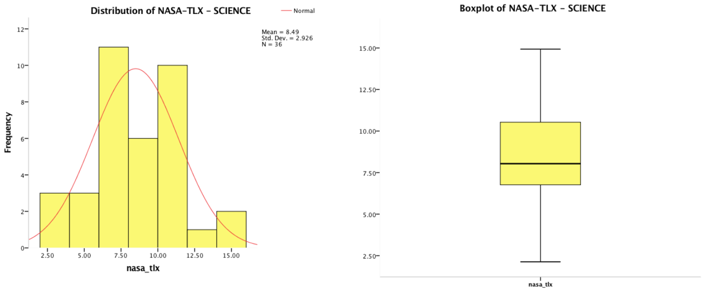
The modelling part was aimed to determine the importance of a keyword that is characteristic of a core text when compared to a contrastive corpus. It was measured as the relative frequency of a keyword in the core text (FT), divided by the relative frequency of the keyword in the contrastive text (FC) as defined on the equation (2.6) in chapter 2. For the NASA-TLX and WP datasets under four core texts, three instructional designs and a contrastive corpus, this section began with the calculation of FT and FC, where the frequency of each keyword in the corpus was divided by the total number of words of the corpus. During that process, mismatches were avoided using a synonym search, based on the similarity of two words under the WUP path-based similarity measure, and the task that was carried out using algorithm 10 proposed in chapter 3. Secondly, each possible synonym and its WUP similarity value was compared based on the maximum value of WUP to obtain the synonym most similar to the keyword using algorithm 11 (see chapter 3). Then, the Relatives frequencies FT and FC were determined using algorithm 12 proposed in chapter 3. For each keyword, a Relative Frequency Ratio (RFR) was calculated. Finally, the average of the Relative Frequency Ratios (RFRavg) for each participant was obtained. Thus, those tasks were achieved using algorithm 13 proposed during the experiment design and methodology.

4.3.1 NASA-TLX

1. **Science**

   (a) **NASA-TLX**

   The distribution of NASA-TLX for the Science topic is analysed based on figure 4.11 along with the result obtained from the test of normality of Kolmogorov-Smirnov shown in table 4.14.



(a) Histogram.

(b) Boxplot.

Figure 4.11: NASA-TLX: Assessment of normality.

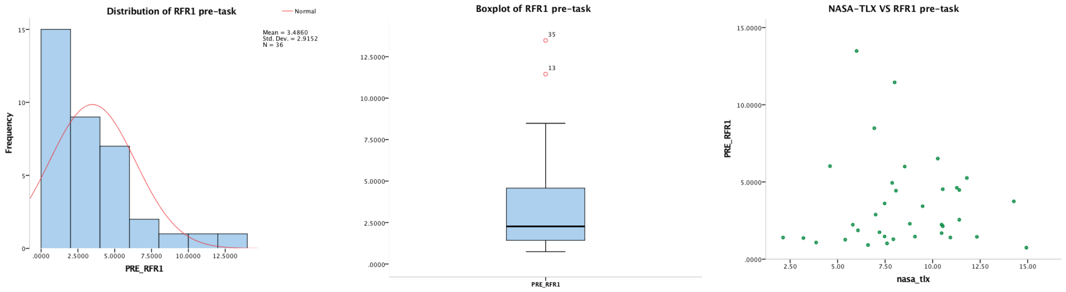
Table 4.14: NASA-TLX: Test of normality.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
NASA-TLX	0.090	36	0.200

Figure 4.11a indicates that for 36 respondents, NASA-TLX has a mean of 8.49 and a standard deviation of 2.926 with a frequency distribution that suggests normality. Furthermore, figure 4.11b indicates the non-presence of outliers. Finally, based on the non-significant result obtained from Kolmogorov-Smirnov, 0.2 > 0.05, it can be said that NASA-TLX for the Science topic is normally distributed.

(b) **Pre-task**

Firstly, the distribution, linearity and homoscedasticity of the Relative Frequency Ratios of keywords under the topic Science pre-task are analysed based on figure 4.12 along with the result obtained from the test of normality of Kolmogorov-Smirnov shown in table 4.15.



(a) Histogram.

(b) Boxplot.

(c) Scatterplot.

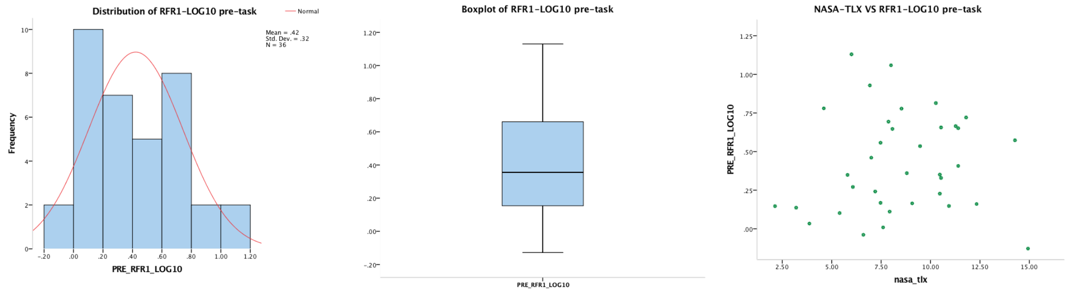
Figure 4.12: Science (RFR1): Assessment of normality, linearity and homoscedasticity pre-task.

Table 4.15: Science (RFR1): Test of normality pre-task.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR1	0.187	36	0.003

Figure 4.12a suggests that for 36 participants, RFR1 pre-task, has a mean of 3.486 and a standard deviation of 2.915 with a skewed to-the-right frequency distribution that indicates non-normality. Furthermore, figure 4.12b suggests the presence of outliers. Based on the significant result obtained from Kolmogorov-Smirnov, $0.003 < 0.05$, it was said that NASA-TLX is not normally distributed. Finally, figure 4.12c shows that there is no indication of a linear relationship between NASA-TLX and RFR1. Also, it is evident that the relationship between both variables is not the same across all values.

Based on the analysis conducted above, the Relative Frequency Ratios of keywords under the topic Science pre-task was transformed applying the $log10$ to it. Thus, the new frequency distribution, linearity and homoscedasticity are analysed based on figure 4.13 along with the result obtained from the second test of normality of Kolmogorov-Smirnov shown in table 4.16.



(a) Histogram.

(b) Boxplot.

(c) Scatterplot.

Figure 4.13: Science (RFR1): Assessment of normality, linearity and homoscedasticity pre-task (2nd test).

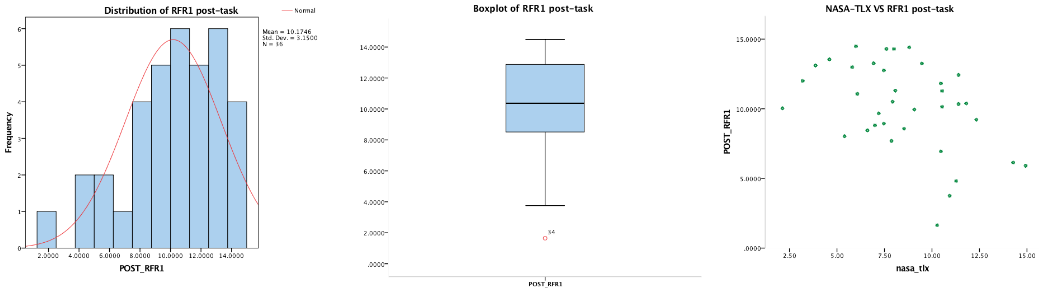
Table 4.16: Science (RFR1): Test of normality pre-task (2nd test).

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR1-LOG10	0.120	36	0.200

Figure 4.13a indicates that for 36 respondents, RFR1 pre-task transformed, has a mean of 0.42 and a standard deviation of 0.32 with a frequency distribution that suggests normality. Furthermore, figure 4.13b indicates the non presence of outliers. Based on the non-significant result obtained from Kolmogorov-Smirnov, 0.2 > 0.05, it can be said that the transformed RFR1 pre-task is normally distributed. Finally, figure 4.13c shows that there is a small indication of a linear relationship between NASA-TLX and RFR1. Also, it is evident that the relationship between both variables is tending to be the same across all values.

(c) **Post-task**

The distribution, linearity and homoscedasticity of the Relative Frequency Ratios of keywords under the topic Science post-task are analysed based on figure 4.14 along with the result obtained from the test of normality of Kolmogorov-Smirnov shown in table 4.17.



(a) Histogram.

(b) Boxplot.

(c) Sccatterplot.

Figure 4.14: Science (RFR1): Assessment of normality, linearity and homoscedasticity post-task.

Table 4.17: Science (RFR1): Test of normality post-task.

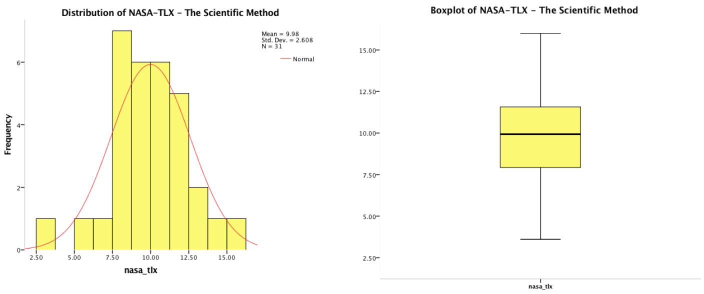
	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR1	0.085	36	0.200

Figure 4.14a indicates that for 36 respondents, RFR1 post-task, has a mean of 10.174 and a standard deviation of 3.15 with a frequency distribution tending to skew to the left. Furthermore, figure 4.14b indicates the presence of one outlier ($ID = 34$). Based on the non-significant result obtained from Kolmogorov-Smirnov, $0.2 > 0.05$, it can be said that RFR1 post-task is normally distributed. From checking the outliers score in the dataset, it was found that the keywords that the participant ($ID = 34$) wrote during the post-task were the least important to Science corpus. Even though it was a genuine score, it was decided to exclude this outlier from the calculations related to it in further analysis. Finally, figure 4.14c shows that there is an indication of a linear relationship between NASA-TLX and RFR1. Also, it is evident that the relationship between both variables tends to be the same across all values.

2. **The Scientific Method**

   (a) **NASA-TLX**

   The distribution of NASA-TLX for the Scientific Method topic is analysed based on figure 4.15 along with the result obtained from the test of normality of Kolmogorov-Smirnov shown in table 4.18.



(a) Histogram.

(b) Boxplot.

Figure 4.15: NASA-TLX: Assessment of normality.

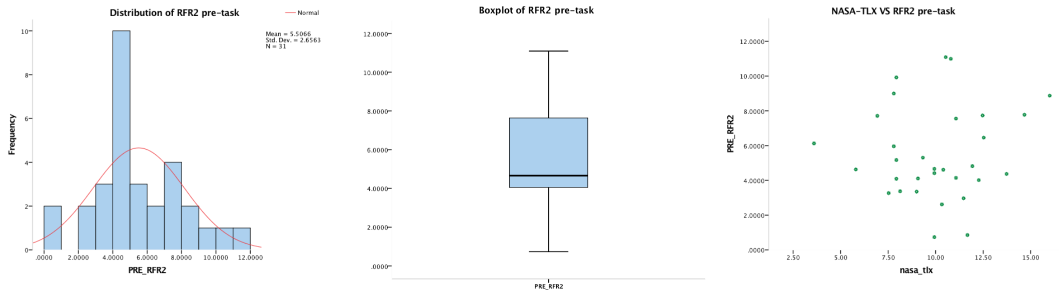
Table 4.18: NASA-TLX: Test of normality.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
NASA-TLX	0.084	31	0.200

Figure 4.15a indicates that for 31 participants, NASA-TLX has a mean of 9.98 and a standard deviation of 2.608 with a frequency distribution that suggest normality. Furthermore, figure 4.15b indicates the non-presence of outliers. Finally, based on the non-significant result obtained from Kolmogorov-Smirnov, 0.2 > 0.05, it can be said that NASA-TLX for the Scientific Method topic is normally distributed.

(b) **Pre-task**

The distribution, linearity and homoscedasticity of the Relative Frequency Ratios of keywords under the Scientific Method Topic are analysed based on figure 4.16 along with the result obtained from applying the test of normality of Kolmogorov-Smirnov shown in table 4.19.



(a) Histogram.

(b) Boxplot.

(c) Scatterplot.

Figure 4.16: The Scientific Method (RFR2): Assessment of normality, linearity and homoscedasticity pre-task.

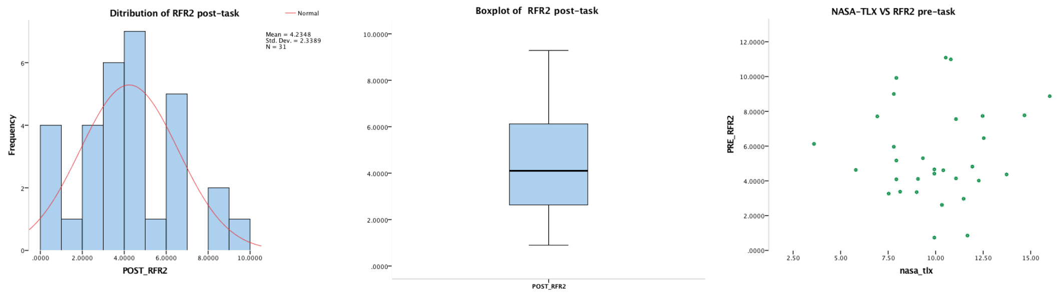
Table 4.19: The Scientific Method (RFR2): Test of normality pre-task.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR2	0.150	31	0.072

Figure 4.16a indicates that for 31 respondents, RFR2 pre-task has a mean of 5.506 and a standard deviation of 2.656 with a frequency distribution that tends to be normally distributed. Furthermore, figure 4.16b indicates the non-presence of outliers. Based on the non-significant result obtained from Kolmogorov-Smirnov, 0.072 > 0.05, it was said that RFR2 pre-task has a normal distribution. Finally, figure 4.16c shows that there is no indication of a linear relationship between NASA-TLX and RFR2. Also, it is evident that the relationship between both variables is not the same across all values.

(c) **Post-task**

The distribution, linearity and homoscedasticity of the Relative Frequency Ratios of keywords under the Scientific Method topic post-task are analysed based on figure 4.17 along with the result obtained from the test of normality of Kolmogorov-Smirnov shown in table 4.20.



(a) Histogram.

(b) Boxplot.

(c) Scatterplot.

Figure 4.17: The Scientific Method (RFR2): Assessment of normality, linearity and homoscedasticity post-task.

Table 4.20: The Scientific Method (RFR2): Test of normality post-task.

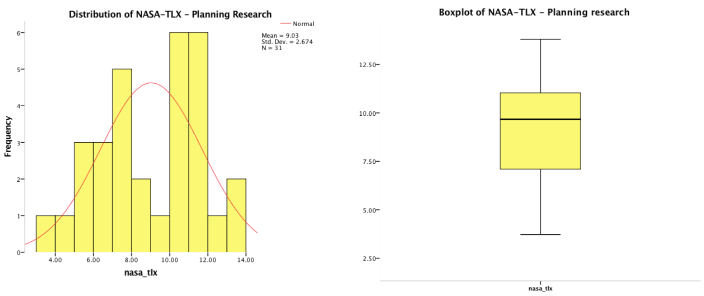
	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR2	0.122	31	0.200

As shown in figure 4.17a, RFR2 post-task, has a mean of 4.234 and a standard deviation of 2.338 with a frequency distribution that tends to be normally distributed. Also, figure 4.17b, indicates the non-presence of outliers. Based on the non-significant result obtained from table 4.20, $0.200 > 0.05$, it can be said that RFR2 post-task has a normal distribution. Finally, figure 4.17a shows that there is no indication of a linear relationship between NASA-TLX and RFR2. Also, it is evident that the relationship between both variables is not the same across all values.

3. **Planning Research**

   (a) **NASA-TLX**

The distribution of NASA-TLX for the Planning Research topic is analysed based on figure 4.18 along with the result obtained from the test of normality of Kolmogorov-Smirnov shown in table 4.21.



(a) Histogram.

(b) Boxplot.

Figure 4.18: NASA-TLX: Assessment of normality.

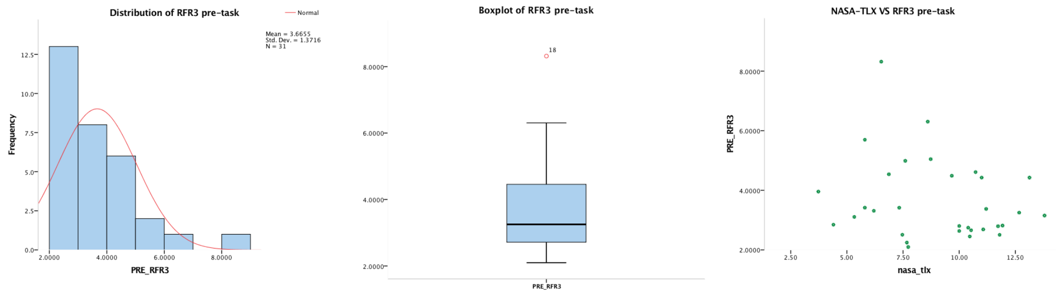
Table 4.21: NASA-TLX: Test of normality.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
NASA-TLX	0.126	31	0.200

Figure 4.18a indicates that for 31 participants, NASA-TLX has a mean of 9.03 and a standard deviation of 2.674 with a frequency distribution that suggest normality. Furthermore, figure 4.18b indicates the non-presence of outliers. Finally, based on the non-significant result obtained from Kolmogorov-Smirnov, $0.2 > 0.05$, it can be said that NASA-TLX for the Planning Research topic is normally distributed.

(b) **Pre-task**

Firstly, the distribution, linearity and homoscedasticity of the Relative Frequency Ratios of keywords under the Planning Research topic pre-task are analysed based on figure 4.19 along with the result obtained from the test of normality of Kolmogorov-Smirnov shown in table 4.22.



(a) Histogram.

(b) Boxplot.

(c) Scatterplot.

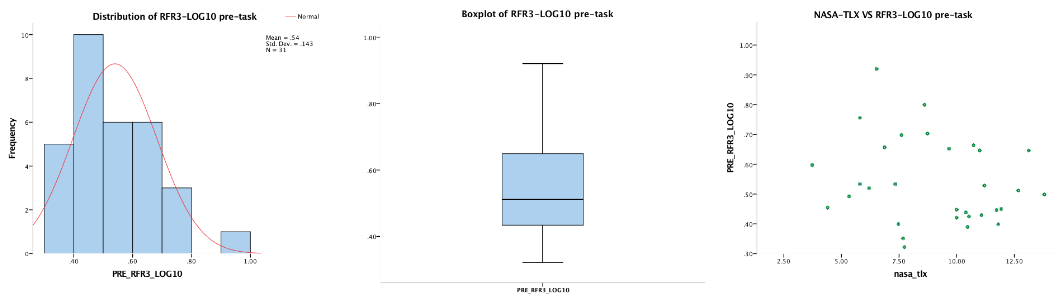
Figure 4.19: Planning Research (RFR3): Assessment of normality, linearity and homoscedasticity pre-task.

Table 4.22: Planning Research (RFR3): Test of normality pre-task.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR3	0.217	31	0.001

Figure 4.19a suggests that for 31 participants, RFR3 pre-task has a mean of 3.665 and a standard deviation of 1.3716 with a skewed to-the-right frequency distribution that indicates non-normality. Furthermore, figure 4.19b suggests the presence of outliers. Based on the significant result obtained from Kolmogorov-Smirnov, $0.001 < 0.05$, it was said that RFR3 pre-task is not normally distributed. Finally, figure 4.19c shows that there is no indication of a linear relationship between NASA-TLX and RFR3. Also, it is demonstrated that the relationship between both variables is not the same across all values.

Based on the analysis conducted above, the Relative Frequency Ratios of keywords under the Planning Research topic pre-task was transformed applying the $log10$ to it. Thus, the new frequency distribution, linearity and homoscedasticity are analysed based on figure 4.20 along with the result obtained from the second test of normality of Kolmogorov-Smirnov shown in table 4.23.



(a) Histogram.

(b) Boxplot.

(c) Scatterplot.

Figure 4.20: Planning Research (RFR3): Assessment of normality, linearity and homoscedasticity pre-task (2nd test).

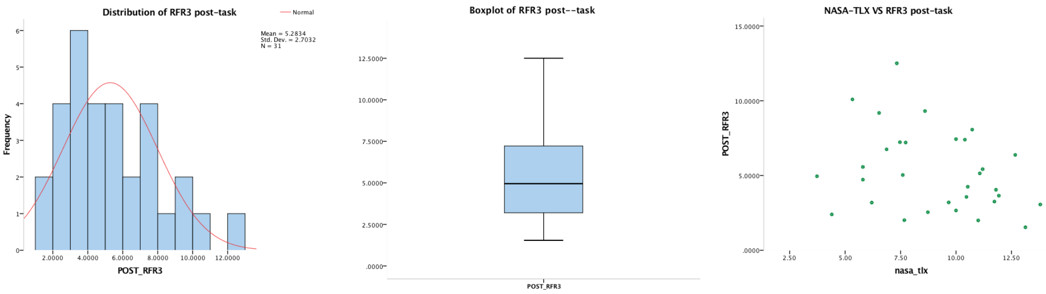
Table 4.23: Planning Research (RFR3): Test of normality pre-task (2nd test).

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR3-LOG10	0.162	31	0.037

Figure 4.20a indicates that for 31 respondents, RFR3 pre-task transformed, has a mean of 0.54 and a standard deviation of 0.143 with a frequency distribution that suggest normality. Furthermore, figure 4.20b indicates the non-presence of outliers. Based on the significant result obtained from Kolmogorov-Smirnov, $0.037 < 0.05$, it can be said that RFR3 pre-task transformed is not normally distributed. Finally, figure 4.20c shows that there is no indication of a linear relationship between NASA-TLX and RFR3. Also, it is shown that the relationship between both variables is not the same across all values.

(c) **Post-task**

The analyses of distribution, linearity and homoscedasticity of the Relative Frequency Ratios of keywords under the Planning research topic post-task are based on figurefig:nasa-post3-t1-h1-b1 and the result obtained from the test of normality of Kolmogorov-Smirnov shown in table 4.24.



(a) Histogram.

(b) Boxplot.

(c) Scatterplot.

Figure 4.21: Planning Research (RFR3): Assessment of normality, linearity and homoscedasticity post-task.

Table 4.24: Planning Research (RFR3): Test of normality post-task.

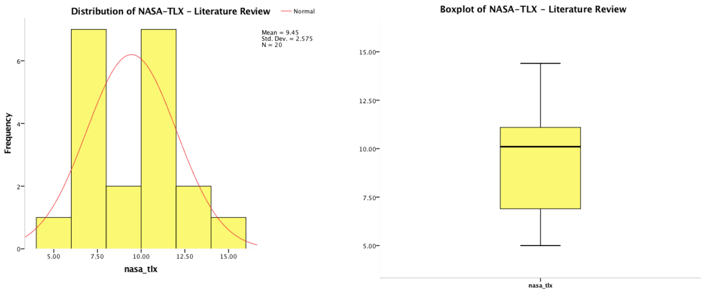
	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR3	0.114	31	0.200

As shown in figure 4.21a, RFR3 post-task has a mean of 5.2834 and a standard deviation of 2.703 with a frequency distribution that suggests normality. Furthermore, figure 4.21b indicates the non-presence of outliers. Based on the non-significant result obtained from table 4.24, 0.2 > 0.05, it can be said that RFR3 post-task is normally distributed. Finally, figure 4.21c shows that there is a small indication of a linear relationship between NASA-TLX and RFR3 that tends to be the same across all values.

4. **Literature Review**

(a) **NASA-TLX**

The distribution of NASA-TLX for the Literature Review topic is analysed based on figure 4.22 and the result obtained from the test of normality of Kolmogorov-Smirnov in table 4.25.



(a) Histogram.

(b) Boxplot.

Figure 4.22: NASA-TLX: Assessment of normality.

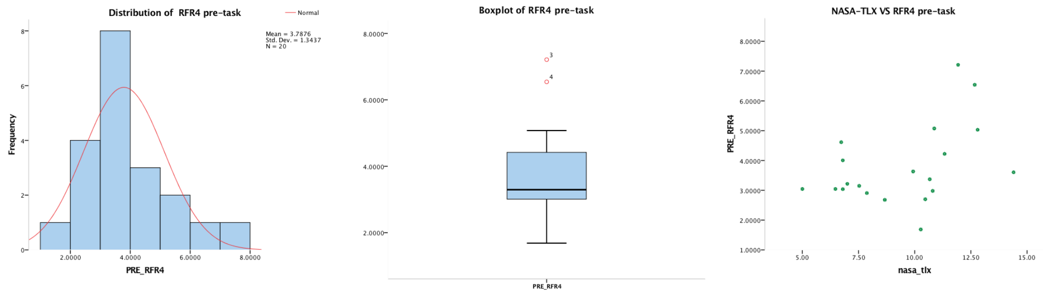
Table 4.25: NASA-TLX: Test of normality.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
NASA-TLX	0.130	20	0.200

Figure 4.22a indicates that for 20 participants, NASA-TLX has a mean of 9.45 and a standard deviation of 2.575 with a frequency distribution that tends to be symmetrical. Furthermore, figure 4.22b indicates the non-presence of outliers. Finally, the non-significant result obtained from table 4.25, $0.2 > 0.05$ suggests that NASA-TLX has a normal distribution.

(b) **Pre-task**

Firstly, the distribution, linearity and homoscedasticity of the Relative Frequency Ratios of keywords under the Literature Review topic pre-task are analysed based on figure 4.23 and the result obtained from the test of normality of Kolmogorov-Smirnov shown in table 4.26.



(a) Histogram.

(b) Boxplot.

(c) Scatterplot.

Figure 4.23: Literature Review (RFR4): Assessment of normality, linearityand homoscedasticity pre-task.

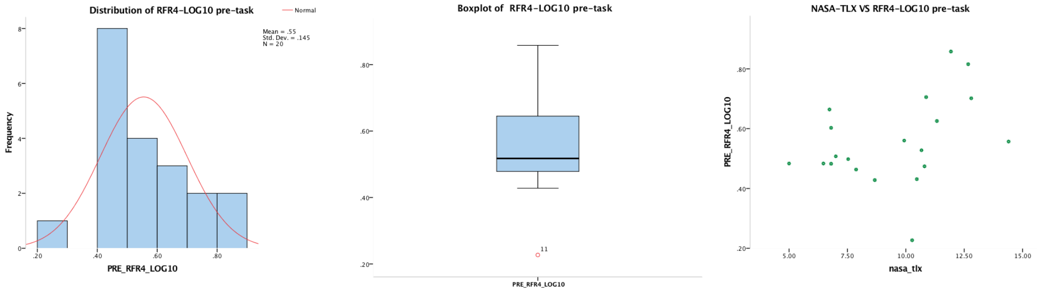
Table 4.26: Literature Review (RFR4): Test of normality pre-task.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR4	0.196	20	0.043

As shown in figure 4.23a, RFR4 pre-task, for 20 participants, has a mean of 3.787 and a standard deviation of 1.344 with a frequency distribution that tends to be normally distributed. Also, figure 4.23b indicates the presence of outliers. The non-significant result obtained from table 4.26, $0.043 < 0.05$, suggests that RFR4 pre-task in not normally distributed. Finally, figure 4.23a shows that there is no indication of a linear relationship between NASA-TLX and RFR4. Also, it is evident that the relationship between both variables is not the same across all values.

Based on the analysis conducted above, the Relative Frequency Ratios of keywords under the Literature Review topic pre-task was transformed applying the $log10$ to it. Thus, the new frequency distribution, linearity and homoscedasticity are analysed based on figure 4.24 and the result obtained from the second test of normality of Kolmogorov-Smirnov is shown in table

4.27.



(a) Histogram.

(b) Boxplot.

(c) Scatterplot.

Figure 4.24: Literature Review (RFR4): Assessment of normality, linearity and homoscedasticity pre-task (2nd test).

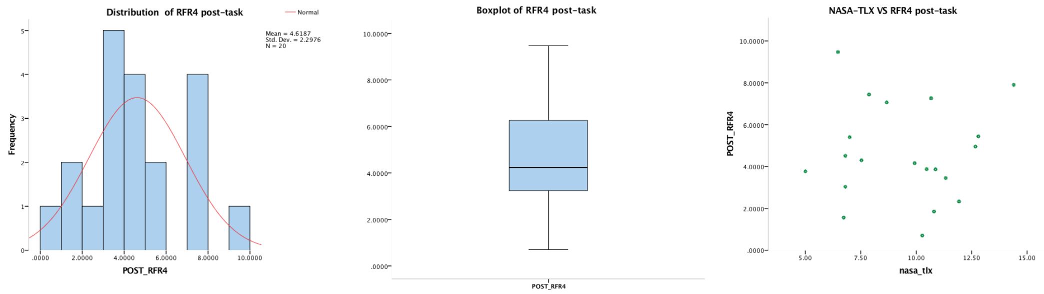
Table 4.27: Literature Review (RFR4): Test of normality pre-task (2nd test).

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR4-LOG10	0.141	20	0.200

As shown in figure 4.24a, the transformed RFR4 pre-task for 20 participants has a mean of 0.55 and a standard deviation of 0.145 with a frequency distribution that tends to be symmetrical. Furthermore, figure 4.24b indicates the presence of one outlier ($ID = 11$). Based on the non-significant result obtained from Kolmogorov-Smirnov, $0.2 > 0.05$, it can be said that the transformed RFR4 pre-task has a normal distribution. From checking the outliers score in the dataset, it was found that the keywords that the participant ($ID = 11$) wrote during the pre-task were the least important to the Literature Review corpus being a genuine score. In this sense, it was decided to exclude this outlier from the calculations related to it in further analysis. Finally, figure 4.24c shows that there is no indication of a linear relationship between NASA-TLX and RFR4. Also, it is shown that the relationship between both variables is not the same across all values.

(c) **Post-task**

The analysis of distribution, linearity and homoscedasticity of the Relative Frequency Ratios of keywords under the Literature Review topic post-task are analysed based on figure 4.25 and the result obtained from the test of normality of Kolmogorov-Smirnov shown in table 4.28.



(a) Histogram.

(b) Boxplot.

(c) Scatterplot.

Figure 4.25: Literature Review (RFR4): Assessment of normality, linearity and homoscedasticity post-task.

Table 4.28: Literature Review (RFR4): Test of normality pre-task (2nd test).

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR4	0.119	20	0.200

Figure 4.25a indicates that, RFR4 post-task, for 20 participants, has a mean of 4.618 and a standard deviation of 2.297 with a frequency distribution that suggests normality. Furthermore, figure 4.25b indicates the non-presence of outliers. The non-significant result obtained from table 4.28, 0.2 > 0.05, suggests that RFR4 post-task has a normal distribution. Finally, figure 4.25c shows that there is no indication of a linear relationship between

NASA-TLX and RFR4. Also, it is demonstrated that the relationship between both variables is not the same across all values.

During this part, 15 tests were conducted for the assessment of normality of NASA-TLX per each topic, namely: Science, The Scientific Method, Planning Research and Literature Review, and also each topic for Relative Frequency Ratios of pre-tasks and post-tasks. It involved the analysis of frequency distributions (histograms), boxplots, scatterplots and $p - values$ (from the test of normality of Kolmogorov-Smirnov) for each variable. In cases where normality was not achieved during a first analysis, a second group of tests were applied after the transformation of the variable using the $log10$ property (based on the skewed to-the-right distribution as in most cases). In this sense, the summary of the results of the 15 tests is presented in table 4.29.

Table 4.29: NASA-TLX & Core texts: table of $p - value$ of Kolmogorov-Smirnov

Kolmogorov-Smirnov (sig-val)						
MWL	**test1**	**Topics**	**Pre-task**		**Post-task**	
			test1	**test2**	**test1**	**test2**
NASA-TLX1	0.2*	Science (n = 36)	0.003	0.2*	0.2*	NA
NASA-TLX2	0.2*	The Scientific Method (n = 31)	0.072*	NA	0.2*	NA
NASA-TLX3	0.2*	Planning Research (n = 31)	0.001	0.037	0.2*	NA
NASA-TLX4	0.2*	Literature Review (n = 20)	0.043	0.2*	0.2*	NA

NA: Not Appplied

*. Variable has a normal distribution.

n: Number of participants

As part of the modelling implementation, the normality of Relative frequency Ratios of keywords under three Instructional Designs (Traditional, Video-delivery and Video-collaborative) were also tested per each corpus (Science, The Scientific Method, Planning research and Literature Review). Thus, 42 tests were conducted for the assessment of normality based on the test of Kolmogorov-Smirnov for each variable. In cases where normality was not achieved during a first analysis, as well as in the previous analysis, a second group of tests were applied. In this sense, the results of these experiments are presented in table 4.30.

Table 4.30: NASA-TLX & Instructional Designs: table of $p - value$ of Kolmogorov-Smirnov

Kolmogorov-Smirnov (sig-val)

MWL	**test1**	**Instructional**	**Topics**	**Pre-task**		**Post-task**	
				test1	**test2**	**test1**	**test2**
NASA-TLX1	0.200*	Traditional (n = 64)	Science	0.000	0.006	0.000	0.010
			The Scientific Method	0.003	0.200*	0.026	0.200*
			Planning Research	0.059*	NA	0.000	0.200*
			Literature Review	0.004	0.200*	0.000	0.200*
NASA-TLX2	0.200*	Video-delivery (n = 44)	Science	0.001	0.169*	0.002	0.000
			The Scientific Method	0.160*	NA	0.000	0..000
			Planning Research	0.000	0.027	0.000	0.013
			Literature Review	0.200*	NA	0.200*	NA
NASA-TLX3	0.200*	Video-Collaborative (n = 9)	Science	0.200*	NA	0.009	0.021
			The Scientific Method	0.200*	NA	0.045	0..200*
			Planning Research	0.200*	NA	0.038	0.079*
			Literature Review	0.200*	NA	0.058*	NA

NA: Not Appplied

*. Variable has a normal distribution.

n: Number of participants

Finally, the results presented above will be used to select the variables that fulfilled or were proximate to achieve normality to calculate the relationship between MWL and RFR.

4.3.2 WP

For the WP dataset, the procedure applied during the previous section was followed as part of the modelling implementation. Thus, 16 tests were conducted for the assessment of normality of WP per each topic, namely: Science, The Scientific Method, Planning Research and Literature Review, and also each topic for Relative Frequency Ratios of pre-tasks and post-tasks. It involved the analysis of frequency distributions (histograms), boxplots and $p - values$ (from the test of normality of Kolmogorov-Smirnov) for each variable. In cases where normality was not achieved during a first analysis, as well as for NASA-TLX dataset, a second group of tests were applied after the transformation of the variable using the $log10$ property (based on the skewed to-the-right distribution as in most cases). The graphs and tables of the tests are annexed in the additional content section (A.0.1). The summary of the results of the 15 tests is presented in table 4.31.

Table 4.31: WP & Core texts: table of $p - value$ of Kolmogorov-Smirnov

Kolmogorov-Smirnov (sig-val)						
MWL	**test1**	**Topics**	**Pre-task**		**Post-task**	
			test1	**test2**	**test1**	**test2**
WP1	0.200*	Science (n = 27)	0.031	0.200*	0.200*	NA
WP2	0.200*	The Scientific Method (n = 29)	0.200*	NA	0.097*	NA
WP3	0.200*	Planning Research (n = 28)	0.002	0.200*	0.010	0.200*
WP4	0.200*	Literature Review (n = 20)	0.003	0.200*	0.200*	NA

NA: Not Appplied

*. Variable has a normal distribution.

n: Number of participants

The normality of Relative frequency Ratios of keywords under three Instructional Designs (Traditional, Video-delivery and Video-collaborative) were also tested per each corpus (Science, The Scientific Method, Planning research and Literature Review). Thus, 43 tests were conducted for the assessment of normality based on the test of Kolmogorov-Smirnov for each variable. In cases where normality was not achieved during a first analysis, as well as in the previous analysis, a second group of tests were applied. The results of these experiments are presented in table 4.32.

Table 4.32: WP & Instructional Designs: table of $p - value$ of Kolmogorov-Smirnov

Kolmogorov-Smirnov (sig-val)

MWL	**test1**	**Instructional**	**Topics**	**Pre-task**		**Post-task**	
				test1	**test2**	**test1**	**test2**
WP1	0.200*	Traditional (n = 67)	Science	0.000	0.196*	0.000	0.012
			The Scientific Method	0.005	0.200*	0.000	0.052*
			Planning Research	0.001	0.200*	0.000	0.039
			Literature Review	0.089*	NA	0.000	0.094*
WP2	0.200*	Video-delivery (n = 28)	Science	0.000	0.129*	0.003	0.088*
			The Scientific Method	0.154*	NA	0.018	0.014
			Planning Research	0.200*	NA	0.021	0.200*
			Literature Review	0.145*	NA	0.049	0.024
WP3	0.200*	Video-Collaborative (n = 9)	Science	0.002	0.058*	0.001	0.020
			The Scientific Method	0.200*	NA	0..200*	NA
			Planning Research	0.200*	NA	0.008	0.200*
			Literature Review	0.005	0.134*	0.200*	NA

NA: Not Appplied

*. Variable has a normal distribution.

n: Number of participants

The results presented above will be used to select the variables that fulfilled or were proximate to achieve normality to calculate the relationship between MWL and RFR.

Chapter 5

Evaluation

This chapter involves the hypotheses testing and the reflection of strengths and limitations of findings (see figure 5.1) based on the analyses performed during the previous section and the results obtained from the evaluation of the relationship between Mental Workload and Relatives Frequency Ratios of keywords gathered during pre-task and post-task activities in third level sessions for the topics Science, The Scientific Method, Planning Research and Literature Review, and also, for the instructional designs, Traditional, Video-delivery and Video-collaborative.

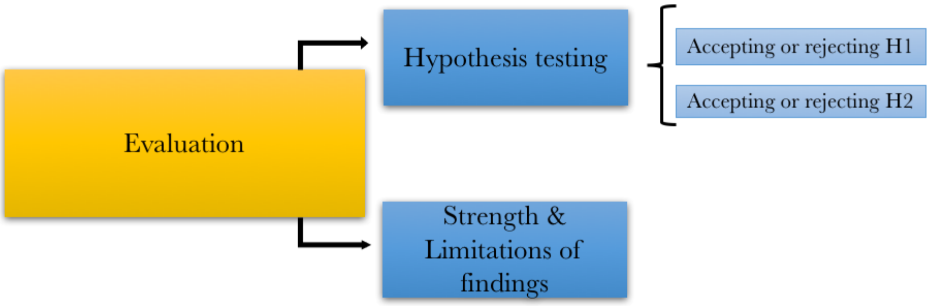

Evaluation
Hypothesis testing
Accepting or rejecting H1
Accepting or rejecting H2
Strength & Limitations of findings

Figure 5.1: Evaluation process

5.1 Hypothesis testing

The nature of the features that are included in this research and the fulfilment of the assumptions of Pearson correlation parametric technique or Spearman correlation non-parametric technique determined which statistical approach was suitable to address the research question. Based on the results obtained during the previous section, the analysis of the relationship between the Mental Workload and Relatives Frequency Ratios is presented as follows:

5.1.1 NASA-TLX

Accepting or rejecting H1

The relationship between NASA-TLX and Relative Frequency Ratios of keywords gathered during MWL pre-task and post-task in third level sessions for the topics Science, The Scientific Method, Planning Research and Literature Review, was investigated using Pearson correlation coefficient (r) and Spearman correlation coefficient (rs). Although preliminary analyses were performed to ensure no violation of the assumptions of normality, linearity, and homoscedasticity for Pearson Correlation, both techniques were applied.

Table 5.1: NASA-TLX & RFR topics (pre-task): Table of correlations.

Comparison	**Topics**	**r**	**p-value**	**rs**	**p-value**
NASA-TLX vs RFR	Science (n = 36)	0.066	0.703	0.148	0.39
	The Scientific Method (n = 31)	0.083	0.656	0.05	0.789
	Planning Research (n = 31)	-0.212	0.251	-0.249	0.177
	Literature Review (n = 20)	0.377	0.101	0.412	0.071

n: Number of participants

Based on the correlations coefficients and $p-values$ shown in table 5.1, it can be stated that there is not a statistically significant relationship between NASA-TLX and RFR of keywords gathered from students during pre-task third-level classes under different topics.

Table 5.2: NASA-TLX & RFR topics (post-task): Table of correlations.

Comparison	Topics	r	p-value	rs	p-value
NASA-TLX vs RFR	Science (n = 36)	-0.446**	0.006	-0.379*	0.023
	The Scientific Method (n = 31)	0.256	0.164	0.245	0.185
	Planning Research (n = 31)	-0.29	0.113	-0.274	0.136
	Literature Review (n = 20)	0.008	0.973	0.058	0.808

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

n: Number of participants

Based on the correlations coefficients and $p - values$ shown in table 5.2, it can be stated that there is a statistically significant relationship between NASA-TLX and RFR of keywords gathered from students during post-task third-level classes under the Science topic with a medium, negative correlation between the two variables, $r = -0.446$, $n = 36$, $p < 0.01$. The two variables that correlate $r = 0.446$ share only 19.89% ($0.446x0.446 = 0.1989 * 100$) of their variance, thus, indicating the presence of an overlap between the two variables. In this sense, the RFR helps to explain nearly 20% of the variance in participants' scores on the Mental Workload scale (NASA-TLX).

Accepting or rejecting H2

The correlation between MWL measures and RFR scores using Video and collaborative (C3), video-delivery approach (C2) and a traditional approach (C1) was investigated using Pearson correlation coefficient (r) and Spearman correlation coefficient (rs) after ensuring no violation of their assumptions.

Table 5.3: NASA-TLX & RFR instructional designs (pre-task): Table of correlations.

Comparison	Topics	r	p-value	rs	p-value
NASA-TLX vs RFR Traditional (n = 64)	Science	-0.011	0.933	0.02	0.874
	The Scientific Method	0.108	0.394	0.087	0.489
	Planning Research	0.126	0.318	0.153	0.224
	Literature Review	-0.119	0.345	-0.103	0.416
NASA-TLX vs RFR Video-Delivery (n = 44)	Science	0.138	0.371	0.206	0.181
	The Scientific Method	0.082	0.599	0.09	0.56
	Planning Research	-0.207	0.177	-0.247	0.106
	Literature Review	-0.141	0.362	-0.157	0.31
NASA-TLX vs RFR Video-collaborative (n = 9)	Science	-0.677	0.065	-0.476	0.233
	The Scientific Method	0.171	0.686	0.333	0.42
	Planning Research	-0.3	0.47	-0.595	0.12
	Literature Review	0.171	0.685	0.405	0.32

n: Number of participants

Based on the Pearson correlation coefficient shown in table 5.3, although the correlations are not statistically significant, H2 is accepted as the assumption $C3 > C2 > C1$ is met when comparing the strength of relationship of MWL and RFR of Instructional designs grouped by Science, Planning and Literature Review topics. On the other hand, H2 is rejected when comparing the strength of relationship of MWL and RFR of the Instructional designs grouped by the Scientific Method topic.

Table 5.4: NASA-TLX & RFR instructional designs: Table of correlations (post-task).

Comparison	Topics	r	p-value	rs	p-value
NASA-TLX vs RFR Traditional (n = 64)	Science	-0.162	0.197	-0.186	0.137
	The Scientific Method	-0.021	0.866	-0.038	0.767
	Planning Research	0.114	0.366	0.12	0.341
	Literature Review	-0.12	0.34	-0.115	0.362
NASA-TLX vs RFR Video-Delivery (n = 44)	Science	-0.033	0.831	0.043	0.783
	The Scientific Method	-0.13	0.4	-0.104	0.503
	Planning Research	-0.288	0.058	-0.282	0.064
	Literature Review	-0.287	0.059	-0.308*	0.042
NASA-TLX vs RFR Video-collaborative (n = 9)	Science	-0.78*	0.023	-0.762*	0.028
	The Scientific Method	0.534	0.173	0.452	0.26
	Planning Research	0.212	0.614	-0.024	0.955
	Literature Review	0.361	0.379	0.286	0.493

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

n: Number of participants

Based on the Pearson correlations coefficient shown in table 5.4, although the correlations are not statistically significant, H2 is accepted as the assumption $C3 > C2 > C1$ is met when comparing the strength of the relationship between MWL and RFR of Instructional designs grouped by The Scientific Method and Literature Review topics. On the other hand, H2 is rejected when comparing the strength of the relationship between MWL and RFR of the Instructional designs grouped by Science and Planning Research topics.

5.1.2 WP

Accepting or rejecting H1

The relationship between WP and Relative Frequency Ratios of keywords gathered during MWL pre-task and post-task in third level sessions for the topics Science, The Scientific Method, Planning Research and Literature Review, was investigated using Pearson correlation coefficient (r) and Spearman correlation coefficient (rs). Although preliminary analyses were performed to ensure no violation of the assumptions of normality, linearity, and homoscedasticity for Pearson Correlation, both techniques were applied.

Table 5.5: WP& RFR topics (pre-task): Table of correlations.

Comparison	**Topics**	**r**	**p-value**	**rs**	**p-value**
WP vs RFR	Science (n = 27)	-0.222	0.267	-0.252	0.205
	The Scientific Method (n = 29)	-0.312	0.099	-0.329	0.082
	Planning Research (n = 28)	0.176	0.37	0.063	0.751
	Literature Review (n = 20)	0.048	0.84	0.055	0.818

n: Number of participants

Based on the correlations coefficients and *pvalues* shown in table 5.5, it can be stated that there is not a statistically significant relationship between WP and RFR of keywords gathered from students during pre-task third-level classes under different topics.

Table 5.6: WP& RFR topics (post-task): Table of correlations.

Comparison	**Topics**	**r**	**p-value**	**rs**	**p-value**
WP vs RFR	Science (n = 27)	-0.091	0.653	-0.136	0.499
	The Scientific Method (n = 29)	0.181	0.346	0.2	0.298
	Planning Research (n = 28)	0.258	0.185	0.347	0.071
	Literature Review (n = 20)	-0.202	0.394	-0.121	0.611

n: Number of participants

Based on the correlations coefficients and *pvalues* shown in table 5.6, it can be stated that there is not a statistically significant relationship between WP and RFR of keywords gathered from students during pre-task third-level classes under different topics.

Accepting or rejecting H2

The correlation between MWL measures and RFR scores using Video and collaborative (C3), video-delivery approach (C2) and a traditional approach (C1) was investigated using Pearson correlation coefficient (r) and Spearman correlation coefficient (rs) after ensuring no violation of their assumptions.

Table 5.7: WP & RFR instructional designs (pre-task): Table of correlations.

Comparison	Topics	r	p-value	rs	p-value
WP vs RFR Traditional (n =67)	Science	-0.264*	0.031	-0.246*	0.045
	The Scientific Method	-0.303*	0.013	-0.274*	0.025
	Planning Research	0.106	0.391	0.155	0.211
	Literature Review	-0.07	0.576	-0.013	0.918
WP vs RFR Video-Delivery (n = 28)	Science	-0.166	0.397	-0.337	0.08
	The Scientific Method	-0.08	0.685	0.025	0.901
	Planning Research	-0.315	0.102	-0.168	0.394
	Literature Review	-0.397*	0.036	-0.251	0.197
WP vs RFR Video-collaborative (n = 9)	Science	0.626	0.071	0.462	0.21
	The Scientific Method	0.478	0.193	0.454	0.22
	Planning Research	-0.159	0.682	-0.042	0.915
	Literature Review	0.57	0.109	0.613	0.079

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

n: Number of participants

Based on the Pearson correlations coefficient shown in table 5.7 although the correlations are not statistically significant, H2 is accepted as the assumption $C3 > C2 > C1$ is met when comparing the strength of relationship between MWL and RFR of Instructional designs grouped by the Literature Review topic. On the other hand, H2 is rejected when comparing the strength of the relationship between MWL and RFR of the Instructional designs grouped by Science, the Scientific Method and Planning Research topics.

Table 5.8: WP & RFR instructional designs (post-task): Table of correlations.

Comparison	Topics	r	p-value	rs	p-value
WP vs RFR Traditional (n =67)	Science	0.169	0.171	0.137	0.27
	The Scientific Method	-0.02	0.872	-0.075	0.544
	Planning Research	0.022	0.859	0.011	0.929
	Literature Review	0.13	0.296	0.116	0.348
WP vs RFR Video-Delivery (n = 28)	Science	-0.162	0.409	-0.155	0.431
	The Scientific Method	-0.117	0.554	-0.097	0.623
	Planning Research	0.141	0.475	0.144	0.464
	Literature Review	-0.134	0.496	-0.054	0.786
WP vs RFR Video-collaborative (n = 9)	Science	0.005	0.99	0.176	0.65
	The Scientific Method	-0.233	0.547	-0.21	0.587
	Planning Research	0.563	0.115	0.65	0.084
	Literature Review	-0.03	0.94	0.286	0.456

n: Number of participants

Based on the Pearson correlations coefficient shown in table 5.8 although the correlations are not statistically significant, H2 is accepted as the assumption $C3 > C2 > C1$ is met when comparing the strength of relationship between MWL and RFR of Instructional designs grouped by The Scientific Method and Planning Research topics. On the other hand, H2 is rejected when comparing the strength of relationship between MWL and RFR of the Instructional designs grouped by Science and Literature Review topics.

5.2 Strengths and limitations of findings

5.2.1 Strengths

The procedures and steps taken to conduct the experiments were clearly identified to achieve high standard results. This research focused specifically on understanding and preparing the data to solve issues that could have affected the interpretation and analysis of text using Natural Language Processing techniques. It solved a major weakness presented during the limitation of designed approach as blank spaces of keywords (missing values) were imputed with the label 'unknown' and handled as stop-words during the data preparation section.

The selection of the statistical techniques was successfully assessed based on the nature of the features that are included in this research and the assumption of the Pearson correlation parametric technique and the Spearman Correlation non-parametric technique. Thus, in the cases where the scores were positively or negatively skewed, transformations were conducted using mathematical formulas making the variables more 'normally' distributed. Although preliminary analyses were performed to ensure no violation of the assumptions of normality, linearity, and homoscedasticity for Pearson Correlation, both techniques were applied.

As the experiments were based on two datasets, NASA-TLX and WP, where each dataset has four topics and three instructional designs, **the combination of the topics and instructional designs facilitated a moderately wide range of analyses** which show that the relationship between Mental Workload and Relative Frequency Ratios of keywords, is only medium correlated, or not correlated at all. Furthermore, from the analyses of multiple cases, it has been found that instructional designs based on the process of hearing and seeing, and the interaction between participants can outperform other approaches, such as those that make use of video supported with images and text, or of a lecturer's speech supported with slides.

5.2.2 Limitations

As identified in the section on the limitations of the designed approach, the first major weakness of this research was **the relatively small datasets which might render it difficult to generate statistically significant results.** Based on that constraint, moderate correlations did not reach statistical significance at the statistical level of $p < 0.05$ which led in more cases to the rejection of the hypothesis H1. In this sense, the quality of the results would be greatly increased if more data is collected.

Another limitation of the results is related to the missing keywords. Although null values of keywords were successfully handled, those imputed missing values might have affected the scores of the Relative Frequency Ratios, decreasing them, which could have influenced the final results when determining the relationship between MWL and RFR.

Finally, another limitation of the research is related to the **misspelling inspection and correction of the text data**, which although was performed using a property of Microsoft Excel because of its easy application, it could have been assessed through the use of a designed algorithm. However, the time was allocated and spent on the identification and replacement of abbreviations, synonyms, word disambiguation in terms of verbs, nouns, adjectives and adverbs and for the calculation of the Mental Workload measures and Relative Frequency Ratios.

Chapter 6

Conclusion

6.1 Research Overview

This research involved five chapters namely, Literature review and related work, Experiment design and methodology, Implementation and results, Evaluation and Conclusion. An overview of their contents is presented as follows:

1. **Chapter 2** outlined a literature review, critically describing related works and the gaps in the fundamentals, namely: Instructional Design, Mental Workload and Natural Language Processing, because they were necessary to formulate the research question. Firstly, the Instructional Design section was presented, which began with the Cognitive Load Theory, answering the question as to when it was developed, its definition and purpose, factors and the relationship between it and instructional designs. Then, it covered types of instructions, namely diverse media and auditory learning, their importance and the related approaches. Then, the Mental Workload section was presented which outlined its foundations including the concept, applications and the negative impact of mental overload and mental underload in performance. Then, the main categories of the Mental Workload measures were presented, followed by the subjective measures NASA Task Load Index and Workload Profile. The Natural Language Processing section began with the factors that have influenced the development of Natural

Language Processing during the last ten years. Then, its definition was presented and the different procedures and applications commonly used. Also, it contextualised the techniques related to the approaches and forms and the research question, namely techniques for text preprocessing, similarity measures and weighting scheme for words. Finally, the related work and summary sections presented the existing work based on Instructional Designs, Mental Workload and Natural Language Processing and the gaps that motivated the formulation of the research question.

2. **Chapter 3** provided a definition of the hypotheses necessary to answer the research question. It also involved software selection, data understanding, data preparation, model design, evaluation and hypotheses testing and strengths and limitations of the designed approach. The first section began with the context that permitted the formulation of the hypotheses that aimed to answer the research question. Then, the hypotheses definition was conducted. The software section involved the selection criteria of the tools that were used to conduct each part of the experiments. Thus, the software were presented along with their tasks to be performed. The data understanding section aimed to identify data quality problems and to discover insights from the data. It involved a number of approaches chosen to analyse the NASA-TLX and WP datasets and also the analysis and explanation of four core texts, namely: Science, The Scientific method, Planning Research, and Literature Review. Finally, the analysis and selection of a contrastive corpus was conducted. For the data understanding of the corpus, two pseudo codes were proposed. The data preparation section presented the steps necessary to solve data quality problems, namely: missing values, outliers, abbreviations, misspellings and assessment of normality during feature generation of the subjective measures of Mental Workload. Also, the steps for feature reduction of the set of keywords features of NASA-TLX and Workload Profile datasets were presented. Finally, during this section five pseudo codes were proposed. The modelling part aimed to determine the importance of a keyword that is characteristic of a core text when compared to

a contrastive corpus. It involved the calculation of the similarity between two words, a synonym search, the calculation of Relative Frequency of keywords in a corpus. Finally, the steps to calculate the Relative Frequency Ratios of keywords were presented. During this section, four pseudo codes were proposed. The evaluation and hypothesis testing section presented the selection of the statistical techniques most suitable to address the research question. Then, different possible scenarios were analysed to accept or reject the hypotheses. Finally, the last section presented the strengths and limitations of the designed approach.

3. **Chapter 4** presented the results of the performed data understanding, data preparation and modelling of the designed research. The first section began with the process of inspecting the NASA-TLX and Workload Profile datasets and five corpus, namely: Science, The Scientific method, Planning Research, Literature Review and Contrastive. It aimed to identify data quality problems and to discover insights into the data, which involved a number of approaches chosen in chapter 3. The data preparation involved the generation of the subjective measures of Mental Workload; the reduction of the set of keywords features of NASA-TLX and Workload Profile datasets; and the data processing of the corpus. Data quality problems, namely: missing values, outliers, abbreviations, stop-words and punctuation symbols, were handled along with the assessment of normality of the variables. Finally, the modelling part aimed to determine the importance of a keyword that is characteristic of a core text when compared to a contrastive corpus. For the NASA-TLX and WP datasets under four core texts, three instructional designs and a contrastive corpus, this section began with the calculation of relative frequencies of keywords. Then, mismatches were avoided using a synonym search, based on the similarity of two words under the WUP path-based similarity measure. Thus, a Relative Frequency Ratio (RFR) was calculated for each keyword and the average of the Relative Frequency Ratios (RFRavg) for each participant was obtained. Finally, the analysis of normality, linearity and homoscedasticity between Mental Workload and Relative Frequency Ratios were conducted.

4. **Chapter 5** involved the hypotheses testing and the reflection of strengths and limitations of findings. It was based on the analyses performed during chapter 4 and the results obtained from the evaluation of the relationship between Mental Workload and Relatives Frequency Ratios of keywords gathered during pre-task and post-task activities in third level sessions for the topics Science, The Scientific Method, Planning Research and Literature Review, and also, for the instructional designs, Traditional, Video-delivery and Video-collaborative.

5. **Chapter 6** concluded by presenting the problems encountered in this thesis. It also discussed the nature of the findings. Moreover, it critically analysed the contribution and impact of this thesis and outlined the implications for future research.

6.2 Problem Definition

To answer the research question, this thesis focused on the following objectives:

1. Investigate the Cognitive Load theory and its relation with instructional designs.

2. Investigate instructional designs and the benefits of a good instructional design to Cognitive Load.

3. Investigate Mental Workload foundations and methods.

4. Investigate Natural Language Processing techniques to analyse importance of keywords in a corpus.

5. Select techniques for synonym search and semantic similarity between two words.

6. Select software according to the most suitable task.

7. Select statistical techniques identifying assumptions.

8. Analyse datasets' features and corpus.

9. Identify data quality problems such as missing values, outliers and abbreviations.

10. Generate and reduce features for the implementation of the model aimed to answer the research question.

6.3 Design/Experimentation, Evaluation & Results

The content of instructional designs translated into text, along with keywords of activities given under those instrument designs, were analysed to measure how those keywords were related to the Mental Workload. In this sense, the first hypothesis was based on the assumption that Relative Frequency Ratios obtained from the keywords collected during the experiments of MWL at third level sessions and the topics in form of text data provide insights of the MWL activity. Thus, the relationship between the variables was investigated using Pearson correlation coefficient (r) and Spearman correlation coefficient (rs). Although preliminary analyses were performed to ensure no violation of the assumptions of normality, linearity, and homoscedasticity for Pearson Correlation, both techniques were applied. Based on that, it was shown that from four corpus, only one gave an indication of a medium, negative relationship between Mental Workload (NASA-TLX) and Relative Frequency Ratios of keywords gathered from students during *post-task* in third-level sessions that helps to explain nearly 20% of the variance in participants' scores on the NASA-TLX scale.

The second hypothesis was an extension of the first assumption but it focused on the instructional designs in terms of the acquisition of learning. In this sense, the hypothesis H2 assumed that the strength of the correlation between MWL measures and RFR scores using Video-collaborative was greater than the video-delivery approach which was greater than when using the traditional approach. Based on that, it was demonstrated that video-collaborative evinced a greater correlation than the other instructional designs in most of the evaluated cases. Thus, it was shown that the process of hearing and seeing, and the interaction between the participants, can promote an optimal germane cognitive load outperforming the video-delivery and the traditional approaches. At the same time, the video-delivery approach outperformed the traditional class, thus indicating that a video of the lecturer supported with images and text instead of a lecturer's speech supported with slides, resulted in greater engagement with the participants.

Finally, diverse forms of instructional designs approaches promote better understanding and enhance the germane cognitive load. An instructional design, properly developed and planned, will have a positive impact on individuals learning and might reflect insights in relation to Mental Workload.

6.4 Contributions and impact

This thesis is a novel research project which made use of algorithms that were a robust integration of Natural Language Processing techniques to analyse Mental Workload. Driven by gaps in the existing literature, this research is a demonstration of the application of new technologies to contribute to the analysis of theoretical approaches. As such, it is an expansion of the existing analysis of framing scholarships to contribute to the evaluation of instructional designs and Mental Workload to learning.

6.5 Future Work & recommendations

It is recommended that this work be extended, applying Supervised Machine Learning models to the processed data which was generated from the data preparation.

The main constraint faced by this research was the limitation of available data from which the experiments were based. This limitation may, however, be overcome if more Mental Workload activities can be conducted to generate more data for the dataset.

References

Allones, J., Martinez, D., & Taboada, M. (2014). Automated mapping of clinical terms into snomed-ct. an application to codify procedures in pathology. *Journal of medical systems*, *38*(10), 134. Retrieved from https://www.researchgate.net/profile/Jose_Allones/publication/265298388_Automated_Mapping_of_Clinical_Terms_into_SNOMED-CT_An_Application_to_Codify_Procedures_in_Pathology/links/540ae8690cf2f2b29a2cdabf/Automated-Mapping-of-Clinical-Terms-into-SNOMED-CT-An-Application-to-Codify-Procedures-in-Pathology.pdf

Boele-Vos, M., Commandeur, J., & Twisk, D. (2017). Effect of physical effort on mental workload of cyclists in real traffic in relation to age and use of pedelecs. *Accident Analysis & Prevention*, *105*(Supplement C), 84 - 94. Retrieved from http://www.sciencedirect.com/science/article/pii/S0001457516304389 (Improving cyclist safety through scientific research, ICSC2015.) doi: https://doi.org/10.1016/j.aap.2016.11.025

Boyer, É. O., Bevilacqua, F., Susini, P., & Hanneton, S. (2017, Mar 01). Investigating three types of continuous auditory feedback in visuo-manual tracking. *Experimental Brain Research*, *235*(3), 691–701. Retrieved from https://doi.org/10.1007/s00221-016-4827-x doi: 10.1007/s00221-016-4827-x

Budanitsky, A., & Hirst, G. (2006, March). Evaluating wordnet-based measures of lexical semantic relatedness. *Comput. Linguist.*, *32*(1), 13–47. Retrieved from http://dx.doi.org/10.1162/coli.2006.32.1.13 doi: 10.1162/coli.2006.32.1.13

Caviedes, J. E., & Cimino, J. J. (2004). Towards the development of a conceptual distance metric for the umls. *Journal of Biomedical Informatics*, *37*(2), 77 - 85. Retrieved from http://www.sciencedirect.com/science/article/pii/S1532046404000218 doi: https://doi.org/10.1016/j.jbi.2004.02.001

Caywood, M. S., Roberts, D. M., Colombe, J. B., Greenwald, H. S., & Weiland, M. Z. (2017). Gaussian process regression for predictive but interpretable machine learning models: An example of predicting mental workload across tasks. *Frontiers in human neuroscience*, *10*, 647. doi: 10.3389/fnhum.2016.00647

Chen, S.-J., Kang, Y.-Y., & Lin, C.-L. (2016, Dec 01). Ergonomic evaluation of video game playing. *Journal of Ambient Intelligence and Humanized Computing*, *7*(6), 845–853. Retrieved from https://doi.org/10.1007/s12652-016-0386-z doi: 10.1007/s12652-016-0386-z

Cinaz, B., Arnrich, B., La Marca, R., & Tröster, G. (2013). Monitoring of mental workload levels during an everyday life office-work scenario. *Personal and ubiquitous computing*, *17*(2), 229–239. https://www.researchgate.net/profile/Bert_Arnrich/publication/235834746_Monitoring_of_mental_workload_levels_during_an_everyday_life_office-work_scenario/links/0912f513f4d67cb118000000.pdf.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences . hilsdale. *NJ: Lawrence Earlbaum Associates*, *2*.

Conroy, J. M., Schlesinger, J. D., & O'Leary, D. P. (2006). Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of the coling/acl on main conference poster sessions* (pp. 152–159). Retrieved from https://www.aclweb.org/anthology/P/P06/P06-2.pdf#page=162

Costley, J., & Lange, C. H. (2017). The effects of lecture diversity on germane load. *The International Review of Research in Open and Distributed Learning*, *18*(2). Retrieved from https://ucd.idm.oclc.org/login?url=https://search-proquest-com.ucd.idm.oclc.org/docview/1931605384?accountid=14507

Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D. S. (2014). Analyzing discourse processing using a simple natural language processing tool. *Discourse Processes*, *51*(5-6), 511-534. Retrieved from `http://dx.doi.org/10.1080/0163853X.2014.910723` doi: 10.1080/0163853X.2014.910723

Dai, H.-J., Lai, P.-T., Chang, Y.-C., & Tsai, R. T.-H. (2015). Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization. *Journal of cheminformatics*, *7*(S1), S14. Retrieved from `https://link.springer.com/article/10.1186/1758-2946-7-S1-S14`

Damerau, F. J. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management*, *29*(4), 433–447. doi: https://doi.org/10.1016/0306-4573(93)90039-G

Dindar, M., Kabak Yurdakul, I., & nan Dnmez, F. (2015). Measuring cognitive load in test items: static graphics versus animated graphics. *Journal of Computer Assisted Learning*, *31*(2), 148–161. Retrieved from `http://dx.doi.org/10.1111/jcal.12086` doi: 10.1111/jcal.12086

Doebler, S., Ryan, A., Shortall, S., & Maguire, A. (2017). Informal care-giving and mental ill-health differential relationships by workload, gender, age and area-remoteness in a uk region. *Health & Social Care in the Community*, *25*(3), 987–999. Retrieved from `http://dx.doi.org/10.1111/hsc.12395` doi: 10.1111/hsc.12395

Fallahi, M., Motamedzade, M., Heidarimoghadam, R., Soltanian, A. R., & Miyake, S. (2016). Effects of mental workload on physiological and subjective responses during traffic density monitoring: A field study. *Applied Ergonomics*, *52*(Supplement C), 95 - 103. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0003687015300399` doi: https://doi.org/10.1016/j.apergo.2015.07.009

Faruqui, M., Dodge, J., Jauhar, S. K., Dyer, C., Hovy, E., & Smith, N. A. (2014). Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*. Retrieved from `https://arxiv.org/pdf/1411.4166.pdf`

Furlan, B., Batanovi, V., & Nikoli, B. (2013). Semantic similarity of short texts in languages with a deficient natural language processing support. *Decision Support Systems*, *55*(3), 710 - 719. Retrieved from http://www.sciencedirect.com/science/article/pii/S0167923613000614 doi: https://doi.org/10.1016/j.dss.2013.02.002

Gambhir, M., & Gupta, V. (2017, Jan 01). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, *47*(1), 1–66. Retrieved from https://doi.org/10.1007/s10462-016-9475-9 doi: 10.1007/s10462-016-9475-9

Garg, S., & Kumar, S. (2016, Aug). Josn: Java oriented question-answering system combining semantic web and natural language processing techniques. In *2016 1st india international conference on information processing (iicip)* (p. 1-6). doi: 10.1109/IICIP.2016.7975361

Gmyzin, D. (2017). A comparison of supervised machine learning classification techniques and theory-driven approaches for the prediction of subjective mental workload. https://arrow.dit.ie/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=1107&context=scschcomdis.

Goel, B. (2017). Developments in the field of natural language processing. *International Journal of Advanced Research in Computer Science*, *8*(3).

Guinard, T. (2016). An algorithm for morphological segmentation of esperanto words. *The Prague Bulletin of Mathematical Linguistics*, *105*(1), 63–76. Retrieved from http://dx.doi.org.ucd.idm.oclc.org/10.1515/pralin-2016-0003

Gupta, S., Nenkova, A., & Jurafsky, D. (2007). Measuring importance and query relevance in topic-focused multi-document summarization. In *Proceedings of the 45th annual meeting of the acl on interactive poster and demonstration sessions* (pp. 193–196). Retrieved from http://repository.upenn.edu/cgi/viewcontent.cgi?article=1769&context=cis_papers

Habib, M. S. (2008). *Improving scalability of support vector machines for biomedical named entity recognition.* University of Colorado at Colorado Springs.

Han, P., Shen, S., Wang, D., & Liu, Y. (2012, May). The influence of word normalization in english document clustering. In *2012 ieee international conference on computer science and automation engineering (csae)* (Vol. 2, p. 116-120). doi: 10.1109/CSAE.2012.6272740

Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2017). Semantic similarity from natural language and ontology analysis. *CoRR*, *abs/1704.05295*. Retrieved from http://arxiv.org/abs/1704.05295

Hart, S. G. (2006). Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 50, pp. 904–908). Retrieved from http://apps.usd.edu/coglab/schieber/psyc792/workload/Hart_HFES_20006.pdf

Hurwitz, J., Kaufman, M., & Bowles, A. (2015). *Cognitive computing and big data analytics.* John Wiley & Sons.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (Vol. 3). Pearson London.

Kalyuga, S., & Singh, A.-M. (2016, Dec 01). Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review*, *28*(4), 831–852. Retrieved from https://doi.org/10.1007/s10648-015-9352-0 doi: 10.1007/s10648-015-9352-0

Kettunen, K., Kunttu, T., & Jrvelin, K. (2005). To stem or lemmatize a highly inflectional language in a probabilistic ir environment? *Journal of Documentation*, *61*(4), 476-496. doi: 10.1108/00220410510607480

Khacharem, A., Zoudji, B., & Kalyuga, S. (2015). Expertise reversal for different forms of instructional designs in dynamic visual representations. *British Journal of Educational Technology*, *46*(4), 756–767. Retrieved from http://dx.doi.org/10.1111/bjet.12167 doi: 10.1111/bjet.12167

Khl, T., Scheiter, K., Gerjets, P., & Gemballa, S. (2011). Can differences in learning strategies explain the benefits of learning from static and dynamic visualizations? *Computers & Education*, *56*(1), 176 - 187. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0360131510002290` (Serious Games) doi: https://doi.org/10.1016/j.compedu.2010.08.008

Lassalle, J., Rauffet, P., Leroy, B., Guérin, C., Chauvin, C., Coppin, G., & Saïd, F. (2017, Sep 01). Communication and workload analyses to study the collective work of fighter pilots: the cowork2 method. *Cognition, Technology and Work*, *19*(2), 477–491. Retrieved from `https://doi.org/10.1007/s10111-017-0420-8` doi: 10.1007/s10111-017-0420-8

Leacock, C., & Chodorow, M. (1998). Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, *49*(2), 265–283.

Lee, B., & Cho, K.-H. (2016). Brain-inspired speech segmentation for automatic speech recognition using the speech envelope as a temporal reference. *Scientific reports*, *6*, 37647. Retrieved from `https://www.nature.com/articles/srep37647?WT.feed_name=subjects_auditory-system`

Lehmann, H., & Guenthner, F. (1991). Discourse analysis for a legal expert system. *Computers and the Humanities*, *25*(2), 81–92. doi: 10.1007/BF00124145

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Van Merriënboer, J. J. G. (2013, Dec 01). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, *45*(4), 1058–1072. Retrieved from `https://doi.org/10.3758/s13428-013-0334-1` doi: 10.3758/s13428-013-0334-1

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. (2015a). A defeasible reasoning framework for human mental workload representation and assessment. *Behaviour & Information Technology*, *34*(8), 758–786. Retrieved from https://www.researchgate.net/profile/Luca_Longo/publication/273439176_A_defeasible_reasoning_framework_for_human_mental_workload_representation_and_assessment/links/55cb0ee608aeb975674a62a7.pdf

Longo, L. (2015b). Designing medical interactive systems via assessment of human mental workload. In *Computer-based medical systems (cbms), 2015 ieee 28th international symposium on* (pp. 364–365).

Longo, L. (2016). Mental workload in medicine: Foundations, applications, open problems, challenges and future perspectives. In *Computer-based medical systems (cbms), 2016 ieee 29th international symposium on* (pp. 106–111). Retrieved from https://arrow.dit.ie/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=1189&context=scschcomcon

Longo, L. (2017). Subjective usability, mental workload assessments and their impact on objective human performance. In *Ifip conference on human-computer interaction* (pp. 202–223).

Longo, L., & Barrett, S. (2010). Cognitive effort for multi-agent systems. In *International conference on brain informatics* (pp. 55–66).

Longo, L., & Dondio, P. (2015). On the relationship between perception of usability and subjective mental workload of web interfaces. In *Web intelligence and intelligent agent technology (wi-iat), 2015 ieee/wic/acm in-*

ternational conference on (Vol. 1, pp. 345–352). Retrieved from https://arrow.dit.ie/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=1191&context=scschcomcon

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.

Lossio-Ventura, J. A., Jonquet, C., Roche, M., & Teisseire, M. (2016, Apr 01). Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*, *19*(1), 59–99. Retrieved from https://doi.org/10.1007/s10791-015-9262-2 doi: 10.1007/s10791-015-9262-2

Lv, C., Liu, H., Dong, Y., & Chen, Y. (2016, Sep 01). Corpus based part-of-speech tagging. *International Journal of Speech Technology*, *19*(3), 647–654. Retrieved from https://doi.org/10.1007/s10772-016-9356-2 doi: 10.1007/s10772-016-9356-2

Manke, S. N., & Shivale, N. (2015). A review on: Opinion mining and sentiment analysis based on natural language processing. *International Journal of Computer Applications*, *109*(4). Retrieved from https://pdfs.semanticscholar.org/86e6/ee4001c2b5925edf3031d8b80dacf1dbbd03.pdf

Manning, C. D., Schütze, H., et al. (1999). *Foundations of statistical natural language processing* (Vol. 999). MIT Press.

Mayer, R. (2017). Using multimedia for e-learning. *Journal of Computer Assisted Learning*, *33*(5), 403–423. Retrieved from http://dx.doi.org/10.1111/jcal.12197 (JCAL-16-266.R1) doi: 10.1111/jcal.12197

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist*, *38*(1), 43–52. Retrieved from http://cmapspublic2.ihmc.us/rid=1KXP7KR7M-8N27KG-1FNL/mayer_moreno_2003.pdf

McInnes, B. T., & Pedersen, T. (2013). Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of biomedical* *informatics*, *46*(6), 1116–1124. Retrieved from http://www.sciencedirect.com/science/article/pii/S1532046413001238

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Morrison, J., Watson, G. S., & Morrison, G. R. (2015). Exploring the redundancy effect in print-based instruction containing representations. *British Journal of Educational Technology*, *46*(2), 423–436. Retrieved from http://dx.doi.org/10.1111/bjet.12140 doi: 10.1111/bjet.12140

Moustafa, K., Luz, S., & Longo, L. (2017). Assessment of mental workload: A comparison of machine learning methods and subjective assessment techniques. In *International symposium on human mental workload: Models and applications* (pp. 30–50). https://www.researchgate.net/publication/318145280_Assessment_of_Mental_Workload_A_Comparison_of_Machine_Learning_Methods_and_Subjective_Assessment_Techniques.

Nenkova, A. (2005). Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *Aaai* (Vol. 5, pp. 1436–1441). Retrieved from https://ocs.aaai.org/Papers/AAAI/2005/AAAI05-228.pdf

Nenkova, A., Vanderwende, L., & McKeown, K. (2006). A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *Proceedings of the 29th annual international acm sigir conference on research and development in information retrieval* (pp. 573–580). Retrieved from http://ants.iis.sinica.edu.tw/3BkMJ9lTeWXTSrrvNoKNFDxRm3zFwRR/36/Context%20Sensitive.pdf

Nguyen, H. A., & Al-Mubaid, H. (2006, May). New ontology-based semantic similarity measure for the biomedical domain. In *2006 ieee international conference on granular computing* (p. 623-628). doi: 10.1109/GRC.2006.1635880

Noy, N. F. (2004, December). Semantic integration: A survey of ontology-based approaches. *SIGMOD Rec.*, *33*(4), 65–70. Retrieved from `http://doi.acm.org.ucd.idm.oclc.org/10.1145/1041410.1041421` doi: 10.1145/1041410.1041421

Ott, T., Wu, P., Paullada, A., Mayer, D., Gottlieb, J., & Wall, P. (2016). Athena – a zero-intrusion no contact method for workload detection using linguistics, keyboard dynamics, and computer vision. In C. Stephanidis (Ed.), *Hci international 2016 – posters' extended abstracts: 18th international conference, hci international 2016, toronto, canada, july 17-22, 2016, proceedings, part i* (pp. 226–231). Cham: Springer International Publishing. Retrieved from `https://doi.org/10.1007/978-3-319-40548-3_38` doi: 10.1007/978-3-319-40548-3_38

Paas, F., & Ayres, P. (2014). Cognitive load theory: A broader view on the role of memory in learning and education. *Educational Psychology Review*, *26*(2), 191–195. doi: http://dx.doi.org.ucd.idm.oclc.org/10.1007/s10648-014-9263-5

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist*, *38*(1), 1-4. Retrieved from `https://doi.org/10.1207/S15326985EP3801_1` doi: 10.1207/S15326985EP3801_1

Paas, F., van Gog, T., & Sweller, J. (2010, Jun 01). Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational Psychology Review*, *22*(2), 115–121. Retrieved from `https://doi.org/10.1007/s10648-010-9133-8` doi: 10.1007/s10648-010-9133-8

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian journal of psychology*, *45*(3), 255–287. Retrieved from `http://dx.doi.org.ucd.idm.oclc.org/10.1037/h0084295`

Pallant, J. (2013). *Spss survival manual.* McGraw-Hill Education (UK).

Panda, S. P., & Nayak, A. K. (2016, Mar 01). Automatic speech segmentation in syllable centric speech recognition system. *International Journal of Speech Technol-*

ogy, *19*(1), 9–18. Retrieved from https://doi.org/10.1007/s10772-015-9320-6 doi: 10.1007/s10772-015-9320-6

Paredes-Valverde, M. A., Valencia-Garca, R., Rodrguez-Garca, M. A., Colomo-Palacios, R., & Alor-Hernndez, G. (2016). A semantic-based approach for querying linked data using natural language. *Journal of Information Science*, *42*(6), 851-862. Retrieved from https://doi.org/10.1177/0165551515616311 doi: 10.1177/0165551515616311

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989, Jan). Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, *19*(1), 17-30. doi: 10.1109/21.24528

Rizzo, L., Dondio, P., Delany, S. J., & Longo, L. (2016). Modeling mental workload via rule-based expert system: a comparison with nasa-tlx and workload profile. In *Ifip international conference on artificial intelligence applications and innovations* (pp. 215–229). Retrieved from https://arrow.dit.ie/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=1197&context=scschcomcon

Rizzo, L., & Longo, L. (2017). Representing and inferring mental workload via defeasible reasoning: A comparison with the nasa task load index and the workload profile. In *Ai3@aiia.*

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology*, *53*(1), 61–86. Retrieved from https://s3.amazonaws.com/academia.edu.documents/31893623/Evaluation_of_Subjective_Mental_Workload.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1511786951&Signature=fHadMnCl9jJo%2BatulLFOEk%2FtZOk%3D&response-content-disposition=inline%3B%20filename%3DEvaluation_of_Subjective_Mental_Workload.pdf

Sapena, E., Padró, L., & Turmo, J. (2013, December). A constraint-based hypergraph partitioning approach to coreference resolution. *Comput. Linguist.*, *39*(4), 847–884.

Retrieved from http://dx.doi.org/10.1162/COLI_a_00151 doi: 10.1162/COLI_a_00151

Schilling, J. (2017). In respect to the cognitive load theory: Adjusting instructional guidance with student expertise. *Journal of Allied Health*, *46*(1), 25E–30E. Retrieved from https://ucd.idm.oclc.org/login?url=https://search-proquest-com.ucd.idm.oclc.org/docview/1911583181?accountid=1450

Schmidt-Weigand, F., & Scheiter, K. (2011). The role of spatial descriptions in learning from multimedia. *Computers in Human Behavior*, *27*(1), 22 - 28. Retrieved from http://www.sciencedirect.com/science/article/pii/S0747563210001445 (Current Research Topics in Cognitive Load Theory) doi: https://doi.org/10.1016/j.chb.2010.05.007

Shadiev, R., Huang, Y.-M., & Hwang, J.-P. (2017, Oct 01). Investigating the effectiveness of speech-to-text recognition applications on learning performance, attention, and meditation. *Educational Technology Research and Development*, *65*(5), 1239–1261. Retrieved from https://doi.org/10.1007/s11423-017-9516-3 doi: 10.1007/s11423-017-9516-3

Singh, J., & Gupta, V. (2016, September). Text stemming: Approaches, applications, and challenges. *ACM Comput. Surv.*, *49*(3), 45:1–45:46. Retrieved from http://doi.acm.org.ucd.idm.oclc.org/10.1145/2975608 doi: 10.1145/2975608

Sloan, T. W., & Lewis, D. A. (2014). Lecture capture technology and student performance in an operations management course. *Decision Sciences Journal of Innovative Education*, *12*(4), 339–355. Retrieved from http://dx.doi.org/10.1111/dsji.12041 doi: 10.1111/dsji.12041

Sun, P.-C., & Cheng, H. K. (2007). The design of instructional multimedia in e-learning: A media richness theory-based approach. *Computers & education*, *49*(3), 662–676. Retrieved from http://i-learn.uitm.edu.my/resources/journal/j3.pdf

Sweller, J., Van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998, Sep 01). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. Retrieved from https://doi.org/10.1023/A:1022193728205 doi: 10.1023/A:1022193728205

Taboada, M., Rodriguez, H., Gudivada, R. C., & Martinez, D. (2017). A new synonym-substitution method to enrich the human phenotype ontology. *BMC bioinformatics*, *18*(1), 446. doi: https://doi-org.ucd.idm.oclc.org/10.1186/s12859-017-1858-7

Taieb, M. A. H., Aouicha, M. B., & Hamadou, A. B. (2014). Ontology-based approach for measuring semantic similarity. *Engineering Applications of Artificial Intelligence*, *36*(Supplement C), 238 - 261. Retrieved from http://www.sciencedirect.com/science/article/pii/S0952197614001833 doi: https://doi.org/10.1016/j.engappai.2014.07.015

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, *97*(2), B25 - B34. Retrieved from http://www.sciencedirect.com/science/article/pii/S0010027705000429 doi: https://doi.org/10.1016/j.cognition.2005.01.006

Tsang, P. S., & Velazquez, V. L. (1996). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, *39*(3), 358–381. Retrieved from https://www.researchgate.net/profile/Pamela_Tsang/publication/14354914_Diagnosticity_and_multidimensional_subjective_workload_ratings/links/5573781608ae7536374fd571.pdf

van Genuchten, E., van Hooijdonk, C., Schler, A., & Scheiter, K. (2014). The role of working memory when learning how with multimedia learning material. *Applied Cognitive Psychology*, *28*(3), 327–335. Retrieved from http://dx.doi.org/10.1002/acp.2998 (ACP-13-0008.R2) doi: 10.1002/acp.2998

Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational psychology review*, *17*(2), 147–177.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human factors*, *50*(3), 449–455. Retrieved from `https://www.researchgate.net/profile/Christopher_Wickens/publication/23157812_Multiple_Resources_and_Mental_Workload/links/0deec518144760291a000000/Multiple-Resources-and-Mental-Workload.pdf`

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on association for computational linguistics* (pp. 133–138). Retrieved from `https://arxiv.org/pdf/cmp-lg/9406033.pdf`

Wu, Z., Xu, H., & Lin, T. (2017). A new automated method for evaluating mental workload using handwriting features. *IEICE Transactions on Information and Systems*, *100*(9), 2147–2155. doi: 10.1587/transinf.2016EDP7354

Xie, B., & Salvendy, G. (2000). Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments. *Work & Stress*, *14*(1), 74-99. Retrieved from `https://doi.org/10.1080/026783700417249` doi: 10.1080/026783700417249

Yang, C., Yang, K.-C., & Yuan, H.-C. (2007). Improving the search process through ontology-based adaptive semantic search. *The Electronic Library*, *25*(2), 234–248. Retrieved from `https://ir.nctu.edu.tw/bitstream/11536/14338/1/000246654600009.pdf`

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics*, *58*(1), 1-17. Retrieved from `https://doi.org/10.1080/00140139.2014.956151` (PMID: 25442818) doi: 10.1080/00140139.2014.956151

Yu, P.-T., Wang, B.-Y., & Su, M.-H. (2015). Lecture capture with real-time rearrangement of visual elements: impact on student performance. *Journal of Computer Assisted Learning*, *31*(6), 655–670. Retrieved from `http://dx.doi.org/10.1111/jcal.12109` doi: 10.1111/jcal.12109

Zhang, J., Yin, Z., & Wang, R. (2015). Recognition of mental workload levels under complex human–machine collaboration by using physiological features and adaptive support vector machines. *IEEE Transactions on Human-Machine Systems*, *45*(2), 200–214. doi: 10.1109/THMS.2014.2366914

Zhang, J.-Y., Liu, S.-L., Feng, Q.-M., Gao, J.-Q., & Zhang, Q. (2017). Correlative evaluation of mental and physical workload of laparoscopic surgeons based on surface electromyography and eye-tracking signals. *Scientific Reports*, *7*(1), 11095. Retrieved from `http://dx.doi.org.ucd.idm.oclc.org/10.1038/s41598-017-11584-4`

Zhu, G., & Iglesias, C. A. (2017, Jan). Computing semantic similarity of concepts in knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, *29*(1), 72-85. doi: 10.1109/TKDE.2016.2610428

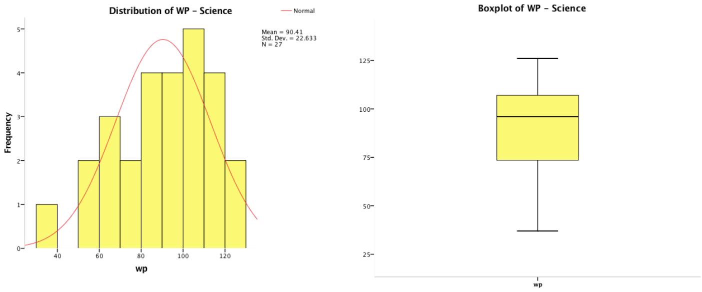
Appendix A

Additional content

A.0.1 WP & RFR: Assessment of normality

1. **Science**

   (a) **WP**



(a) WP: distribution.

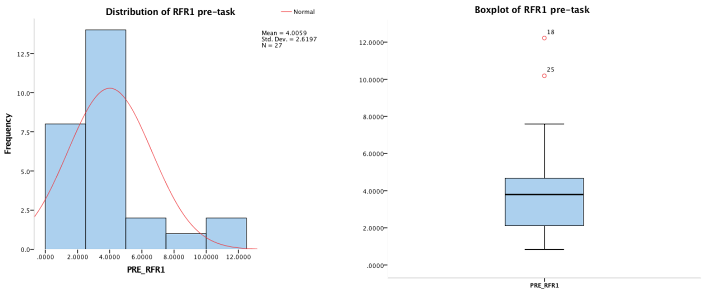
(b) WP: boxplot.

Figure A.1: WP: Assessment of normality.

Table A.1: WP: Test of normality.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
WP	0.116	27	0.200

(b) **Pre-task**



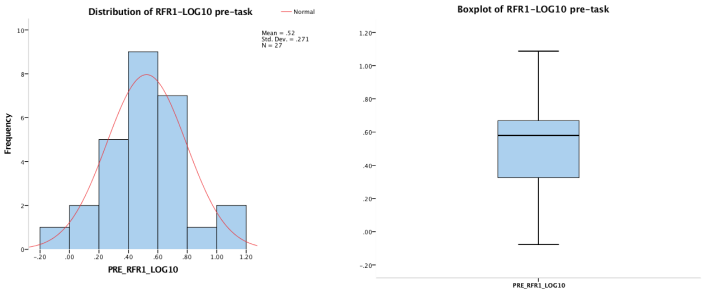
(a) Science (RFR1): distribution pre-task.

(b) Science (RFR1): boxplot pre-task.

Figure A.2: Science (RFR1): Assessment of normality pre-task.

Table A.2: Science (RFR1): Test of normality pre-task.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR1	0.176	27	0.031



(a) Science (RFR1): distribution pre-task (2nd test).

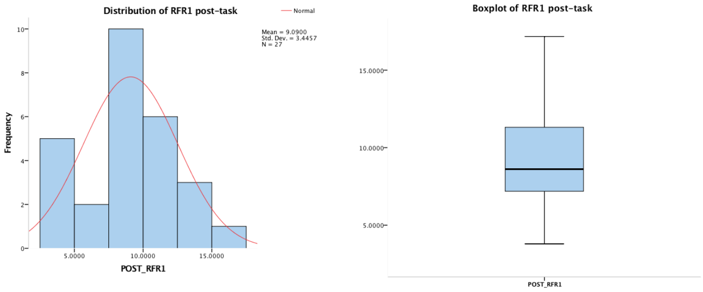
(b) Science (RFR1): boxplot pre-task (2nd test).

Figure A.3: Science (RFR1): Assessment of normality pre-task (2nd test).

Table A.3: Science (RFR1): Test of normality pre-task (2nd test).

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR1-LOG10	0.109	27	0.200

(c) **Post-task**



(a) Science (RFR1): distribution post-task.

(b) Science (RFR1): boxplot post-task.

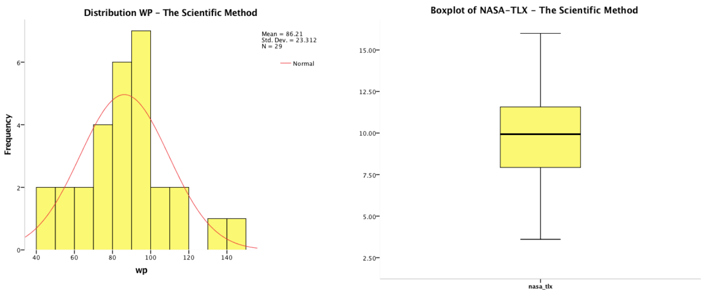
Figure A.4: Science (RFR1): Assessment of normality post-task.

Table A.4: Science (RFR1): Test of normality post-task.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR1	0.085	27	0.200

2. **The Scientific Method**

(a) **WP**



(a) WP: distribution.

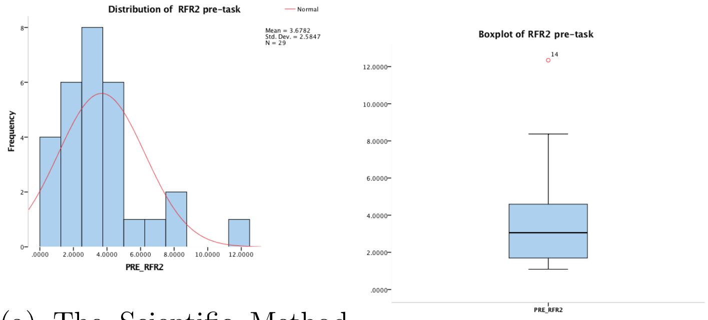
(b) WP: boxplot.

Figure A.5: WP: Assessment of normality.

Table A.5: NWP: Test of normality.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
WP	0.130	29	0.200

(b) **Pre-task**



(a) The Scientific Method (RFR2): distribution pre-task.

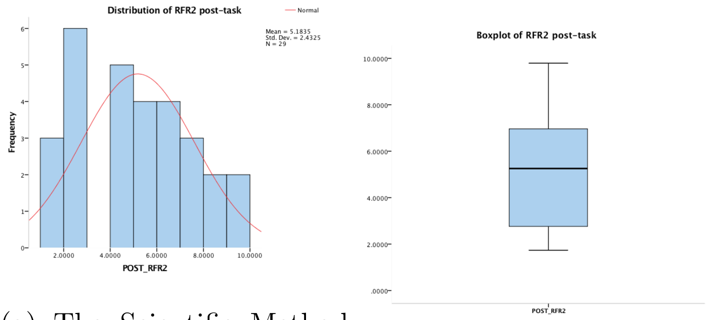
(b) The Scientific Method (RFR2): boxplot pre-task.

Figure A.6: The Scientific Method (RFR2): Assessment of normality pre-task.

Table A.6: The Scientific Method (RFR2): Test of normality pre-task.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR2	0.158	29	0.062

(c) **Post-task**



(a) The Scientific Method (RFR2): distribution pre-task.

(b) The Scientific Method (RFR2): boxplot pre-task.

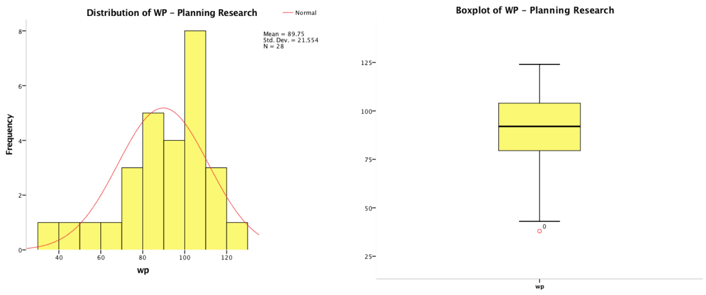
Figure A.7: The Scientific Method (RFR2): Assessment of normality post-task.

Table A.7: The Scientific Method (RFR2): Test of normality post-task.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR2	0.149	29	0.097

3. **Planning Research**

(a) **WP**



(a) WP: distribution.

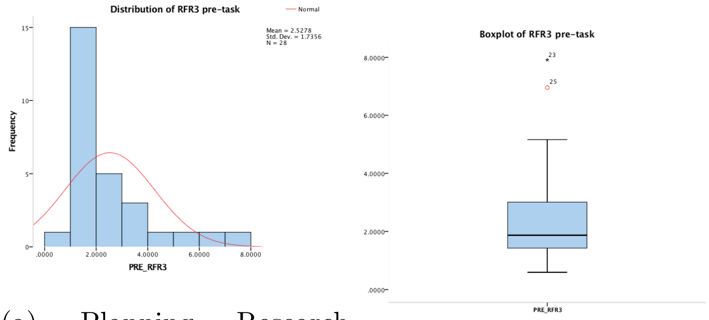
(b) WP: boxplot.

Figure A.8: WP: Assessment of normality.

Table A.8: WP: Test of normality.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
WP	0.128	28	0.200

(b) **Pre-task**



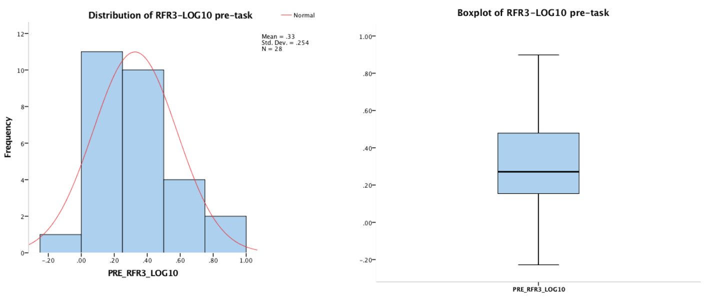
(a) Planning Research (RFR3): distribution pre-task.

(b) Planning Research (RFR3): boxplot pre-task.

Figure A.9: Planning Research (RFR3): Assessment of normality pre-task.

Table A.9: Planning Research (RFR3): Test of normality pre-task.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR3	0.213	28	0.002



(a) Planning Research (RFR3): distribution pre-task (2nd test).

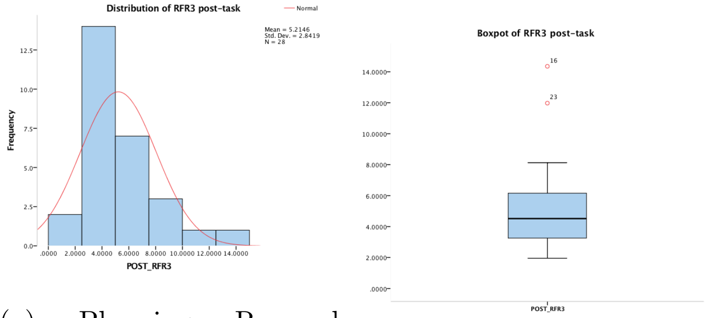
(b) Planning Research (RFR3): boxplot pre-task (2nd test).

Figure A.10: Planning Research (RFR3): Assessment of normality pre-task (2nd test).

Table A.10: Planning Research (RFR3): Test of normality pre-task (2nd test).

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR3-LOG10	0.125	28	0.200

(c) **Post-task**



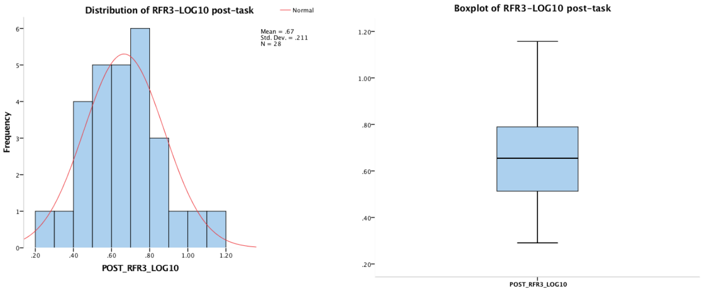
(a) Planning Research (RFR3): distribution post-task.

(b) Planning Research (RFR3): boxplot post-task.

Figure A.11: Planning Research (RFR3): Assessment of normality post-task.

Table A.11: Planning Research (RFR3): Test of normality post-task.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR3	0.191	28	0.010



(a) Planning Research (RFR3): distribution post-task (2nd test).

(b) Planning Research (RFR3): boxplot post-task (2nd test).

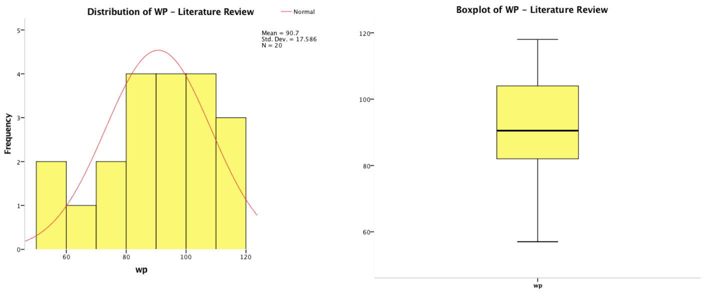
Figure A.12: Planning Research (RFR3): Assessment of normality post-task (2nd test).

Table A.12: Planning Research (RFR3): Test of normality post-task (2nd test).

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR3-LOG10	0.095	28	0.200

4. **Literature Review**

   (a) **WP**



(a) WP: distribution.

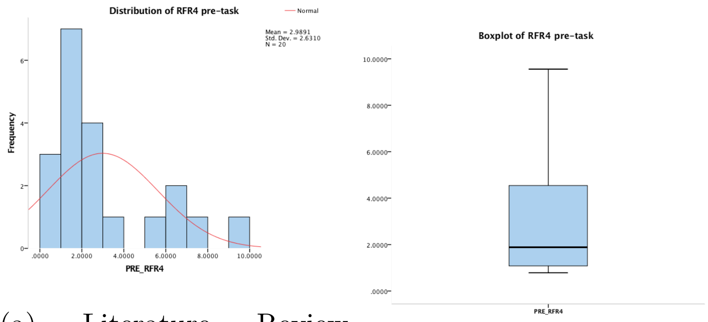
(b) WP: boxplot.

Figure A.13: WP: Assessment of normality.

Table A.13: WP: Test of normality.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
WP	0.123	20	0.200

   (b) **Pre-task**



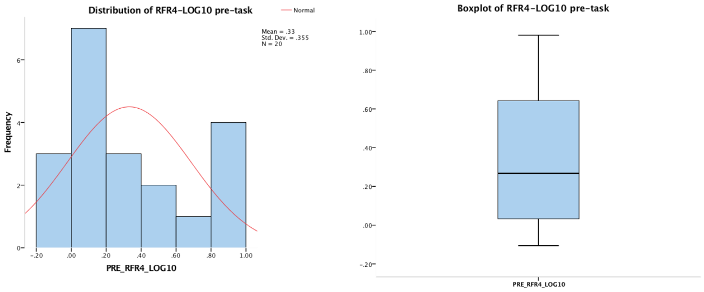
(a) Literature Review (RFR4): distribution pre-task.

(b) Literature Review (RFR4): boxplot pre-task.

Figure A.14: Literature Review (RFR4): Assessment of normality pre-task.

Table A.14: Literature Review (RFR4): Test of normality pre-task.

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR4	0.243	20	0.003



(a) Literature Review (RFR4): distribution pre-task (2nd test).

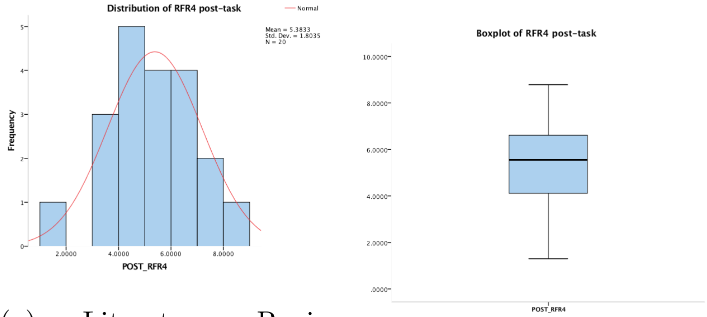
(b) Literature Review (RFR4): boxplot pre-task (2nd test).

Figure A.15: Literature Review (RFR4): Assessment of normality pre-task (2nd test).

Table A.15: Literature Review (RFR4): Test of normality pre-task (2nd test).

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR4-LOG10	0.156	20	0.200

(c) **Post-task**



(a) Literature Review (RFR4): distribution post-task.

(b) Literature Review (RFR4): boxplot post-task.

Figure A.16: Literature Review (RFR4): Assessment of normality post-task.

Table A.16: Literature Review (RFR4): Test of normality pre-task (2nd test).

	Kolmogorov-Smirnov		
	Statistic	**df**	**Sig.**
RFR4	0.099	20	0.200