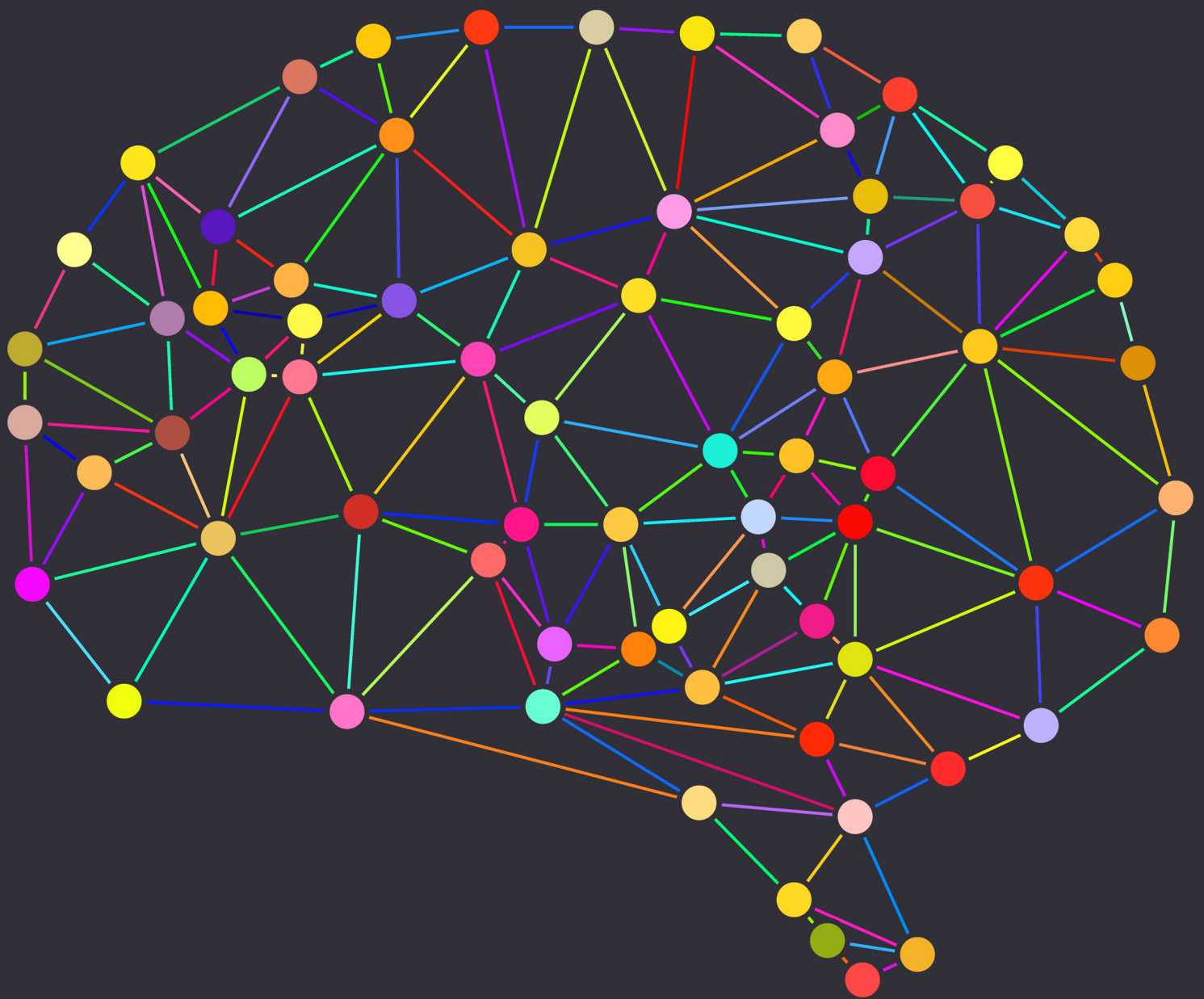


Developing a New Measure of Conceptual Knowledge

# CONCEPT RETRIEVAL TECHNIQUE



Gavin Hays



ISBN: 978-0-646-80127-8

© Gavin Hays, 2019

No part of this thesis may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of the author

**Developing a New Measure of Conceptual Knowledge**

# **The Concept Retrieval Technique**

**Gavin Hays**



**DEVELOPING A NEW MEASURE OF CONCEPTUAL KNOWLEDGE:  
THE CONCEPT RETRIEVAL TECHNIQUE**

Ontwikkeling van een nieuw instrument om conceptuele kennis mee te meten:

De concept retrieval techniek

**Proefschrift**

ter verkrijging van de graad van doctor aan de

Erasmus Universiteit Rotterdam

op gezag van de

Rector Magnificus

Prof. dr. R.C.M.E Engels

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

Vrijdag 3 mei 2019 om 13:30 uur

door

**Gavin James Hays**

geboren te *Sydney, Australie*

**Erasmus University Rotterdam**

The logo of Erasmus University Rotterdam, featuring a stylized, cursive script of the word "Erasmus" in a dark blue or black color.

## **PROMOTIECOMMISSIE**

### **Promotoren:**

Prof. dr. H.G. Schmidt

Prof. dr. H.T. van der Molen

### **Overige leden:**

Prof. dr. L.R. Arends

Prof. dr. F. Paas

Prof. dr. A.A.C.M Smeets

### **Copromotor:**

Dr. J.I. Rotgans

## CONTENTS

VOORWOORD .....	9
<b>CHAPTER 1: GENERAL INTRODUCTION.....</b>	<b>10</b>
1.1 Introduction.....	12
1.2 Background of the problem.....	12
1.3 Statement of the problem .....	14
1.4 Purpose of the study.....	14
1.5 Research questions .....	15
<b>CHAPTER 2: REVIEW OF THE LITERATURE.....</b>	<b>16</b>
2.1 Introduction.....	18
2.2 The Psychological underpinnings of the Concept Retrieval Technique.....	18
2.3 Semantic Network Theory - Hierarchical network model.....	21
2.4 Semantic Network Theory – Spreading activation model .....	25
2.5 Semantic Network Theory – Connectionist model.....	30
2.6 The neuropsychological evidence of Semantic Network Theory .....	33
2.7 The Concept Retrieval Technique and its application to education.....	37
<b>CHAPTER 3: THE RELIABILITY OF THE CONCEPT RETRIEVAL TECHNIQUE.....</b>	<b>41</b>
3.1 Reliability evidence for the Concept Retrieval Technique.....	42
3.2 Study 1: Inter-rater agreement while scoring the Concept Retrieval Technique .....	44
3.3 Study 2. Stability of Inter-rater agreement over subjects, age groups and raters.....	47
3.4 Study 3. Scoring students’ responses in full sentences.....	50
3.5 Summary of findings.....	53
<b>CHAPTER 4: THE VALIDITY OF THE CONCEPT RETRIEVAL TECHNIQUE.....</b>	<b>54</b>
4.1 Validity evidence for the Concept Retrieval Technique.....	56
4.2 Study 4. Convergent validity of the Concept Retrieval Technique.....	57
4.4 Study 5. Construct validity of the Concept Retrieval Technique .....	60
4.5 Summary of findings.....	64
4.6 Discussion of key findings.....	65
<b>CHAPTER 5: DESIGNING AN AUTOMATED CONCEPT RETRIEVAL TECHNIQUE.....</b>	<b>67</b>
5.1 Introduction.....	68
5.2 The nature of automated assessment in an educational context .....	68
5.3 The feasibility of machine-scoring in an educational context .....	69
5.4 The Concept Retrieval Technique as an effective instrument for automated assessment.....	71
5.5 Design considerations for an automated Concept Retrieval Technique.....	74
5.6 The use of modularisation to construct the automated Concept Retrieval Technique.....	75
5.6.1 Version 1 – Reliable scoring engine .....	76
5.6.2 Version 2 – Linguistic analysis.....	77
5.6.3 Version 3 – Import and export data features.....	77

<b>CHAPTER 6: AUTOMATED CONCEPT RETRIEVAL TECHNIQUE (VERSION 1)</b>	<b>79</b>
6.1 Introduction	80
6.2 Design objectives	80
6.2.1 Objective 1 – Test-taker data collection and storage	80
6.2.2 Objective 2 – Creating a dynamic online database	81
6.2.3 Objective 3 – Uploading CSV data files to the online database	81
6.2.4 Objective 4 – Scoring all relevant concepts	82
6.3 Database design	83
6.4 User interface design	84
6.5 System modelling	87
6.5.1 The automated Concept Retrieval Technique (Main Module)	87
6.5.2 Create database (Submodule)	88
6.5.3 Upload data set (Submodule)	89
6.5.4 Scoring engine (Submodule)	90
6.6 Module construction	91
6.6.1 Database connection (Submodule)	91
6.6.2 Upload dataset (Submodule)	93
6.6.3 Saving the target word list (Submodule)	95
6.6.4 Searching for a target concept (Submodule)	97
6.6.5 Scoring engine (Submodule)	100
6.7 Study 6 - Reliability of the automated Concept Retrieval Technique (Version 1)	102
6.8 Summary of key findings	105
<b>CHAPTER 7: AUTOMATING THE SCORING OF THE CONCEPT RETRIEVAL TECHNIQUE (VERSION 2)</b>	<b>108</b>
7.1 Introduction	110
7.2 Design objectives	110
7.2.1 Objective 1 – Improved search and scoring functionality of the scoring engine	111
7.2.2 Objective 2 – Generating word cloud visualisations of test-taker responses	114
7.2.3 Objective 3 – Data cleaning of test-taker responses	115
7.3 System modelling	115
7.3.1 The automated Concept Retrieval Technique (Main Module)	115
7.3.2 Upload dataset (Submodule)	116
7.3.3 Clean dataset (Submodule)	118
7.3.4 Scoring engine (Submodule)	120
7.4 Module construction	120
7.4.1 Upload dataset (Submodule)	120
7.4.2 Clean dataset (Submodule)	123
7.4.3 Scoring engine (Submodule)	125
7.5 Study 7 - Reliability of the automated Concept Retrieval Technique (Version 2)	127
7.6 Study 8 – Generalizability of the Automated Concept Retrieval Technique (Version 2)	130
7.7 Summary of key findings	133
<b>CHAPTER 8: AUTOMATED CONCEPT RETRIEVAL TECHNIQUE (VERSION 3)</b>	<b>135</b>
8.1 Introduction	136
8.2 Design Objectives	136
8.2.1 Objective 1 – Downloading the scored test-taker responses	137
8.2.2 Objective 2 – Enhancing the upload dataset to remove redundant characters	137
8.2.3 Objective 3 – Allowing user interactivity in the construction of the data visualisation	137
8.2.4 Objective 4 – Data cleaning of responses to include spelling recommendations	137
8.2.5 Objective 5 – Improve user interface instructions and other help resources	138

8.2.6	Objective 6 – Movement of the program from a local to an online environment.....	140
8.3	System Modelling.....	140
8.3.1	The automated Concept Retrieval Technique (Main Module).....	140
8.3.2	Download dataset (Submodule).....	142
8.4	Module Construction .....	143
8.4.1	Download dataset (Submodule).....	143
8.4.2	Upload dataset (Submodule).....	145
8.4.3	Clean dataset (Submodule).....	146
8.4.4	Help (Submodule).....	147
8.5	Online Environment for beta testing.....	149
8.6	Study 9 – Stability of the Automated Concept Retrieval Technique (Version 3) .....	150
8.7	Summary of key findings.....	153
<b>CHAPTER 9: SUMMARY AND CONCLUSIONS .....</b>		<b>155</b>
9.1	Introduction.....	156
9.2	Summary of findings.....	156
9.3	Shortcomings .....	160
9.4	Directions for further research.....	162
<b>SAMENVATTING (Summary in Dutch).....</b>		<b>164</b>
<b>REFERENCES.....</b>		<b>175</b>
<b>CURRICULUM VITAE.....</b>		<b>188</b>
<b>AUTHOR PUBLICATIONS.....</b>		<b>192</b>





## VOORWOORD

I have been fortunate to have many kind and brilliant people who have motivated me throughout the completion of this thesis. I am deeply indebted to their support and kindness!

First, I would like to thank my Promotors, Prof Henk van der Molen and Prof Henk Schmidt, who carried the responsibility of this endeavour. I am particularly thankful for their unwavering patience, guidance and constructive feedback during the course of this thesis. The trust that they have bestowed in me to undertake this opportunity on the other side of the world can never be repaid. In addition, I thank them for pushing me at the crucial moments to ensure that I remained motivated till the end.

Second, I would like to thank my Daily Main Supervisor, Dr Jerome Rotgans, who made this thesis possible through his academic guidance on a day-to-day basis. He taught me the ropes from how to design a study to its execution and preparing the final manuscript. He patiently guided me through each step involved in the scientific endeavour and provided invaluable support and encouragement along the way.

In addition, a special thanks to Mabel Tan for her work undertaken in the development of the first two studies in this thesis and a heartfelt appreciation to Adam Hendry for his assistance through the academic writing process and his role as a rater in the manual scoring of the Concept Retrieval Technique. To the Members of the Doctorate Committee: Professor Dr Lidia Arends, Professor Dr Fred Paas and Professor Dr Guss Smeets. Thank you for taking the time and effort to read my thesis. Also, I would like to thank the Catholic Education Diocese of Parramatta and in particular Brother Patrick Howlett, Sue Walsh and Greg Whitby, who without their generous support, this thesis would not have been possible.

Lastly, I would like to thank my family for their unending support and sacrifice in my completion of this thesis. In particular my wife Catherine, who was the backbone of our young family providing me with the faith to pursue this opportunity and the motivation to help me persevere through the tough times. Hopefully, I will be given the opportunity to repay her in the future.







# 1

**Chapter**

**General Introduction**

## 1.1 Introduction

The primary focus of this thesis is to propose a new measure of conceptual knowledge, that can be applied in diverse educational environments. In this chapter, the background and objectives for the thesis will be discussed by first presenting the problem statement and providing more details of the purpose for the proposed study. This will be followed by stating testable research questions. Finally, some limitations to the current research will be highlighted.

## 1.2 Background of the problem

Although many educational reforms have taken place in the past four decades that have changed the way we teach our students, little changes have occurred in how we assess them (Brown, Bull, & Pendlebury, 1997; Glass & Sinha, 2013). The implementation of active-learning pedagogies, has changed the purpose of assessment from summative to formative (Bell & Cowie, 2001; Yorke, 2003), but this does not constitute a change of the assessment format itself, but a change in how the responses are interpreted and used (e.g., MCQs can be used for formative purposes as well). In particular, schools that have adopted active-learning pedagogies, such as problem-based learning, still heavily rely on conventional assessment formats such as multiple-choice (MCQ) or true/false items. Consequently, the narrow knowledge domain assessed by MCQs is thought to promote surface learning, which is often in contradiction with the instructional method, and denies students the opportunity to express a particular depth and scope of knowledge within a subject area (Liu, Lee, & Linn, 2011; Simkin & Kuechler, 2005).

The most commonly used repertoire of assessment formats in education boils down to a handful of test types: multiple-choice, single-best answer, true/false, matching items, fill-in-the-blank, and short answer or essay items (Gronlund, 1998). These test formats can be categorised according to two fundamental cognitive mechanisms: recognition vs. retrieval (Gay, 1980; Jonsson & Svingby, 2007; Nicol, 2007). All test formats, except the latter two (short answer or essay items) can be classified as recognition tasks. For these tests, the test-takers are expected to evaluate all possible answer options, which are provided as cues, and matches the correct answer with knowledge stored in memory. Additionally, the nature of these assessment formats may unintentionally encourage students to select the correct answer by guessing (Nnodim, 1992). In order to increase the reliability of such items, many more items need to be administered, which increases the overall construction and complexity of this assessment instrument (Tarrant, Ware, & Mohammed, 2009; Wass, Van der Vleuten, Shatzer, & Jones, 2001).

Scoring these tests is relatively easy since it can be automated and adequate levels of reliability can be obtained by increasing the number of test items and thus reducing opportunities for guessing the correct answer (Nicol, 2007). Thus, to achieve adequate levels of reliability, these recognition types of tests typically require relatively many items to be administered (Wass, Van der Vleuten, Shatzer, & Jones, 2001). An important disadvantage of recognition-type tests is that, in everyday life the knowledge user has to *retrieve* knowledge from memory rather than to just recognize it. Recognition test have limited authenticity.

In contrast to recognition, retrieval tasks require the test-takers to retrieve knowledge from memory without cues hinting to the correct answer as it is the case for the recognition-type of items. This process is considered more difficult as it requires more cognitive resources to be employed and it is often perceived as a more adequate measure of a student's knowledge and understanding (Bacon, 2003). This is partially due to the fact that guessing is not possible with these tests as compared with recognition tests. What students retrieve and write down, is considered to be a representation of what that person knows about the topic in question.

However, a common operational drawback of using retrieval tests, such as essay-type items, is that they require labour-intensive scoring procedures to ensure acceptable levels of reliability (Jonsson & Svingby, 2007). The process to score an essay-type item requires raters to read through each response. In order to consistently mark the responses provided by the test-takers, a detailed marking scheme is typically required. A marking scheme, sometimes referred to as rubric, is a template that specifies (and exemplifies) what to look out for in a response text to award marks (Jonsson & Svingby, 2007; Reddy & Andrade, 2010). Given each response text needs to be read, interpreted and scored according to the marking scheme, inconsistencies in the reliability of the marks frequently occur. To address these reliability issues, it is common practice to have two markers independently score the responses and determine their agreement, by means of an inter-rater agreement score (Jonsson & Svingby, 2007; Rezaei & Lovorn, 2010; Stellmack, Konheim-Kalkstein, Manor, Massey, & Schmitz, 2009)

Based on the different criteria specified in the scoring rubric, it is then the raters task to interpret and assign the mark accordingly. To achieve reliability in scoring, ideally two independent raters are required to score the responses (Lane, Liu, Ankenmann, & Stone, 1996; Moskal & Leydens, 2000). The scores provided by the two raters can then be compared to examine how far the raters agree with each other in providing the same scores to the same student. A low agreement between raters requires adjustment of the marking scheme and all responses need to be remarked until satisfactory levels of agreement are reached.

In sum, test types can be categorised as either being recognition or retrieval tasks. Retrieving knowledge from memory is considered a more authentic representation of a test-taker's knowledge since no answers are provided and guessing can be eliminated. However, knowledge-retrieval tests require labour-intensive scoring to reach acceptable levels of reliability. To find a trade-off between both types of test formats, an alternative format is required that promotes free recall and yet can reliably be scored.

### **1.3 Statement of the problem**

To find an alternative assessment format that promotes free recall and can be easily scored, a test needs to be designed that is rooted in cognitive principles of how knowledge is organised and retrieved from the human mind. The new measure should address the following criteria: (a) an alternative assessment format that is valid and can be reliably be scored, (b) one that relies on retrieval rather than recognition and (c) utilise automation opportunities to improve the efficiency of the measure, reliability and overall educational utility.

### **1.4 Purpose of the study**

In this study, the Concept Retrieval Technique will be proposed as a viable alternative to overcome the shortcomings of conventional assessment measures. The Concept Retrieval Technique requires students to freely recall concepts or ideas for a given topic. For instance, in a secondary school science topic for “motion”, students would be required to list all the concepts and ideas that they can associate with the trigger – “motion.” The scoring process is by means of a list containing all admissible concepts (the “target word list”), which is typically devised by subject-matter experts beforehand to assure content coverage (or content validity). For instance, in an educational setting, the target word list is based on the key concepts extracted from the learning objectives, the syllabus, and other learning resources students have to study. One would expect to see concepts such as force, friction and Newton in the student responses. Often two independent raters are utilised to assign one mark for each correctly retrieved concept for every respondent. Scoring can be done relatively quickly because the rater only has to focus on the correctness of retrieved concepts and assign one mark for each correct concept. This scoring procedure is in stark contrast to scoring conventional open-ended items that require the rater to read through relative lengthy responses and deciding how many marks to assign.

The retrieval mechanism in the Concept Retrieval Technique is based on spreading activation theory. Basically, this theory states that from one input concept, activation will spread along the links to other concepts that are connected to it. Then, from each of these concepts to

connected others, and so on. The scores for the Concept Retrieval Technique are hypothesized to represent the knowledge network including relevant connections for an individual student. It is developed based on well-established research findings from cognitive psychology on how knowledge is organized in semantic networks of connected concepts. Finally, the automation of the Concept Retrieval Technique will be undertaken in a number of version iterations. The primary aim of the automation will be to maintain consistent inter-rater reliability with human raters to ensure the effectiveness of machine-scoring. Each version of the automated Concept Retrieval Technique will build upon the limitations identified in the testing of the previous version. Finally, the automated Concept Retrieval Technique will be transitioned to an online environment where beta testing can be thoroughly conducted by users.

### **1.5 Research questions**

Following from the above, this thesis will address the following research questions:

1. How reliable is the Concept Retrieval Technique as a new measure of conceptual knowledge to be used consistently across different school subjects and age groups?
2. Given the nature of the Concept Retrieval Technique, how can it be utilised as a valid measure of conceptual understanding?
3. What is the correlation between student performance on the Concept Retrieval Technique and their performance on the conventional assessments for the same topics?
4. What validity evidence can be provided to determine whether the Concept Retrieval Technique measures what it is intended to measure?
5. How reliable is the machine-scoring of the Concept Retrieval Technique in comparison to human scoring?





**Chapter**

**2**

**Literature Review**



## 2.1 Introduction

In this chapter, a literature review will summarise the theoretical underpinnings of the Concept Retrieval Technique and its application to the field of education. An elaboration of established classical and contemporary memory models from cognitive psychology will be presented, supporting the idea that humans capture information about the world from repeated episodic experience, to construct semantic memory networks. In addition, neuropsychological evidence will be presented that can shed more light on how effective retrieval and activation patterns in the brain support the Concept Retrieval Technique. Cognitive scientists have designed experiments to develop theories to provide a more concrete understanding of these mental representations. The network representation is the most relevant representation in explaining how the Concept Retrieval Technique works and will be the primary focus of this literature review. As the Concept Retrieval Technique is proposed to be used in educational settings, the last part of this review will address the use of concept maps as an educational instrument that applies the semantic network theory. Although concept maps may be a good measure of students' knowledge structures, it has some significant limitations with regards to its validity and reliability. While the purpose of the Concept Retrieval Technique appears to be similar to the approach of concept mapping, the methodology of the Concept Retrieval Technique could potentially address the shortcomings faced by concept mapping.

## 2.2 The Psychological underpinnings of the Concept Retrieval Technique

The Concept Retrieval Technique is based on the idea that a representation of a person's knowledge is built from a network of dynamically linked concepts, stored in long-term memory (Champagne, Klopfer, Desena, & Squires, 1981; Collins & Loftus, 1975; Kiefer & Pulvermüller, 2012). Over time, within domain specific learning episodes, students develop richer and more tightly integrated semantic networks of concepts (Glaser & Bassok, 1989; Jonassen, Beissner, & Yacci, 1993; Vinet & Zhedanov, 2011). The purpose of the Concept Retrieval Technique is to assess a student's conceptual knowledge by measuring the number of concepts a student can correctly recall. There is a clear distinction in the literature between the two main types of knowledge in learning: (1) conceptual knowledge and, (2) procedural knowledge. Critical to the validation of the Concept Retrieval Technique is conceptual knowledge, specifically the idea that this knowledge is accumulated as a representation of concepts, often from our sensory and motor experiences (Jones, Willits, & Dennis, 2015; Kiefer & Pulvermüller, 2012).

A plethora of research studies ranging from physics education (Koponen & Pehkonen, 2010) to medical education (Bordage, 1994; Charlin, Tardif, & Boshuizen, 2000; Patel & Groen, 1986) have demonstrated that conceptual knowledge is organized in networks of related concepts (Brachman, 1977; Collins & Loftus, 1975; Collins & Quillian, 1969). The way cognitive psychologists represent conceptual knowledge in the human mind is that concepts are mental representations of what we know about the world, including information about an object, a depiction of an object, or a set of objects indicated verbally (e.g., by a word) (Rogers & McClelland, 2004). For instance, looking at a painting of a *dog* in a park will activate information stored in these mental representations, often based on facts and our experiences. The concept *dog* may activate information such as has *four legs*, a *tail*, *wet noses*, are *mammals* and are used as *pets*. Each concept shares a connection to each other by a linking word, known as a “concept-link-concept chain” or a “proposition” (Kiefer & Pulvermüller, 2012; Roberts & Joiner, 2007). The verification of the statement a “dog has a tail” requires accessing our stored mental representations to determine the statements validity. Although, the verification of whether the *grass* and *leaves* in the painting, are the same colour, does not require access to our mental representations. This judgement can be determined based on the colour information provided in the painting, without reference to stored information about the concepts (Rogers & McClelland, 2004). Figure 2.1 illustrates the concept-link-concept chain with *dog* and *tail* depicted as unique concepts and a connection represented by a linking word (e.g., has).

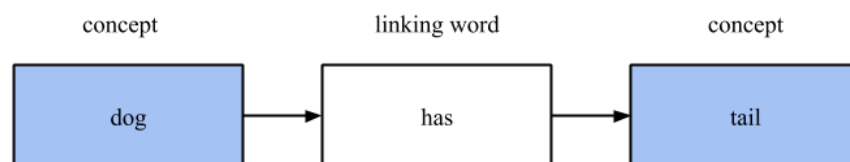
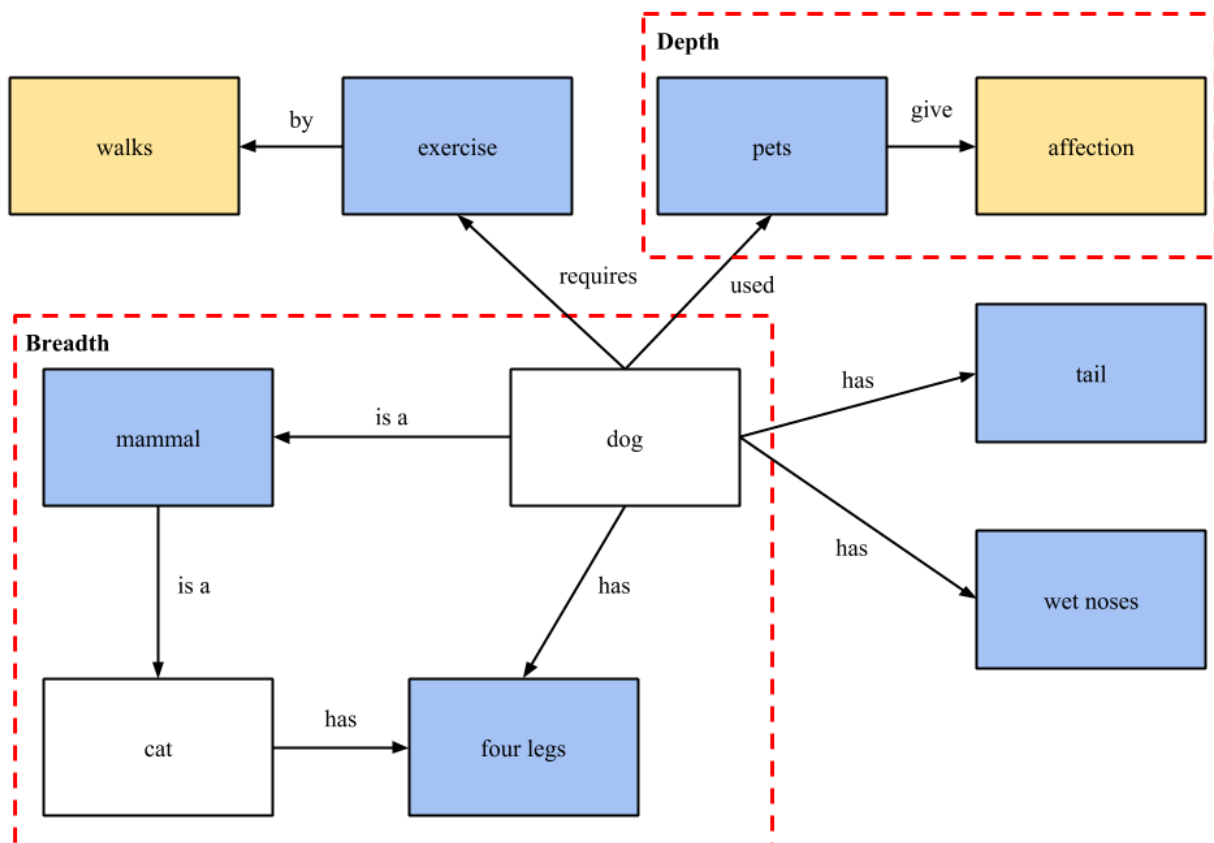


Figure 2.1. An illustration on the representation of the concept dog in its simplest form.

Semantic models are constructed to demonstrate how semantic information is represented and used in cognitive processing. Naturally, human cognition requires access to much more sophisticated and complex structures, exhibiting multiple concept representations, built from linking words to form concept-link-concept chains. Figure 2.2 shows a more complex representation for the concept *dog*. This representation includes the simple preposition from Figure 2.1 but extends the number of concepts that can be linked to *dog*, demonstrating a depth and breadth of the network of conceptual knowledge. Every concept is organised by a set of defining

features. Although these features can exist on their own, they are jointly used to define the concept and all members of the concept can share these features. For instance, an object could only be classified as a *dog*, if it displayed all the listed features. Therefore, if the object did not have a wet nose, it could not be classified as a *dog*. The concepts in the blue boxes represent attribute features of the concept *dog*. However, the interconnectedness of conceptual knowledge shows depth in the yellow boxes for example, *pets* is an attribute of the concept *dog*, but the concept *affection* is an attribute of *pets*. Therefore, the statement “*dogs give affection*” is true, based on the concept-link-chain. In addition, breadth of knowledge is expressed by the connections of the concepts *dog* and *cat*, by means of *mammals* and *four legs*. Hence, the idea that conceptual knowledge is represented as connected webs of concepts is pivotal in helping to explain the use and application of semantic networks.



*Figure 2.1.* An illustration of a more complex representation of the concept *dog*, highlighting breadth and depth of conceptual knowledge.

The idea of semantic networks, first introduced by Ross Quillian (1967), attempted to make sense of human cognition and how memory is an associative structure, developed by information obtained from our experiences with the world, which was later defined as semantic memory (Chang, 1986; Tulving, 1986). The structure of a semantic network consists of concept nodes and labelled preposition links to symbolize the relationship between knowledge representations (Quillian, 1967). In the context of a semantic network, each node holds a unique concept, linked to other concept nodes. The depth of knowledge is represented by the next level of concept nodes (e.g., *pets give affection*), who also share connections to other nodes, a process that continues until all the connections have been exhausted (Raaijmakers & Shiffrin, 1992). The theory of semantic networks states that the entire network representation always begins with the first node representing the full meaning of any concept. See Figure 2.3 in the next section for a detailed example of a semantic network representing the concept *animal*. In addition, the next section will address how researchers have come up with the hierarchical network model to further document the representation and organization of conceptual knowledge. Since shortcomings were later found in the model, a revised model was proposed, which has been used to explain scientific cognitive mechanisms, such as the recall of words, which is important in testing how the Concept Retrieval Technique works.

### 2.3 Semantic Network Theory - Hierarchical network model

The semantic network theory has been one of the most influential theoretical frameworks used to discern how semantic knowledge is represented in the mind. The earliest classical model, known as the hierarchical network model was proposed by Collins and Quillian (1969) and explains both conceptual and propositional knowledge within a single framework (Jones et al., 2015). This model assumes that knowledge is stored in a hierarchical structure with associations between concepts occurring over different levels of hierarchy. These associative links are typically subset-superset prepositions (e.g., *A robin is a bird*) and attribute prepositions (e.g., *A robin has wings*) (Quillian, 1967). It is said in this theory of semantic network that the entire network as entered from the first node represents the full meaning of any concept. For instance, the concept “*bird*” has nodes such as “*a bird can fly*” or “*a bird has wings*” (Collins & Quillian, 1969; Tulving, 1986).

The structure referred to in the above example occurs in the form of a hierarchy. This model describes the most general categories of concepts at the top of the network, while the concept descriptions become more specific towards the bottom of the hierarchy. This model proposed by

Collins and Quillian (1969) outlines that concepts are organised and categorised according to different levels of hierarchy, which can reduce the amount of space required for storage. For example, the three-level hierarchy described in Figure 2.3 represents the possible retrieval paths linked to the concept *animal*. The first level starts with the more general node *animal*, which is categorised by the concept-link-concept chain, has *skin* and can *eat*. As a result, any concept that is linked to *animal*, must adhere to the stated attributes. The second level is composed of *bird* and *fish*, which are subordinate nodes to *animal*. Likewise, the concepts *robin* and *ostrich*, share a subordinate relationship to the superordinate node *bird*. This superordinate and subordinate structure of nodes produces the hierarchical tree structure, which allows the model to explain for conceptual and propositional knowledge within a single framework. Conceptual knowledge is knowledge that is rich in relationships and is often thought of as in the context of a network, while propositional knowledge is more descriptive, which can be expressed in using declarative sentences or propositions.

The processing of this model is undertaken by starting with the highest superordinate node and traversing linked subordinate nodes according to the verification of prepositions statements. Collins and Quillian (1969) used an example of a sentence verification task to demonstrate the activation pathway. The verification of the sentence “*a robin can breathe*”, requires both the subject term *robin* and its property *breathe* to be found in the network. The process will initiate from the node *robin* and the search will continue through its immediate properties, which is *red-breasted* and *fly*. Unfortunately, the property *breathe* is unable to be found within the node *robin*, therefore, the search process accesses the linked superordinate node *bird* and its linked properties. In this instance, the top level of the hierarchy containing the node *animal*, needs to be verified. Given both the subject *bird* and its property *breathe* have been found within the same network, the concept is therefore true. Each time the search process is required to move to higher levels of hierarchy, time is consumed and reduces the overall efficiency of the model. Similarly, if the property cannot be found, the process will continue searching the superordinate nodes and their relevant properties until all nodes have been exhausted, determining the statement as false. In addition, the opportunity for property inheritance and generalization of new knowledge enables automatic inheritance of subordinate concepts, promoting the overall economy of the model (Rogers & McClelland, 2004). For example, adding the concept *perch* as a subordinate of the more general category *fish*, would enable *perch* to automatically inherit the existing knowledge and properties of fish.

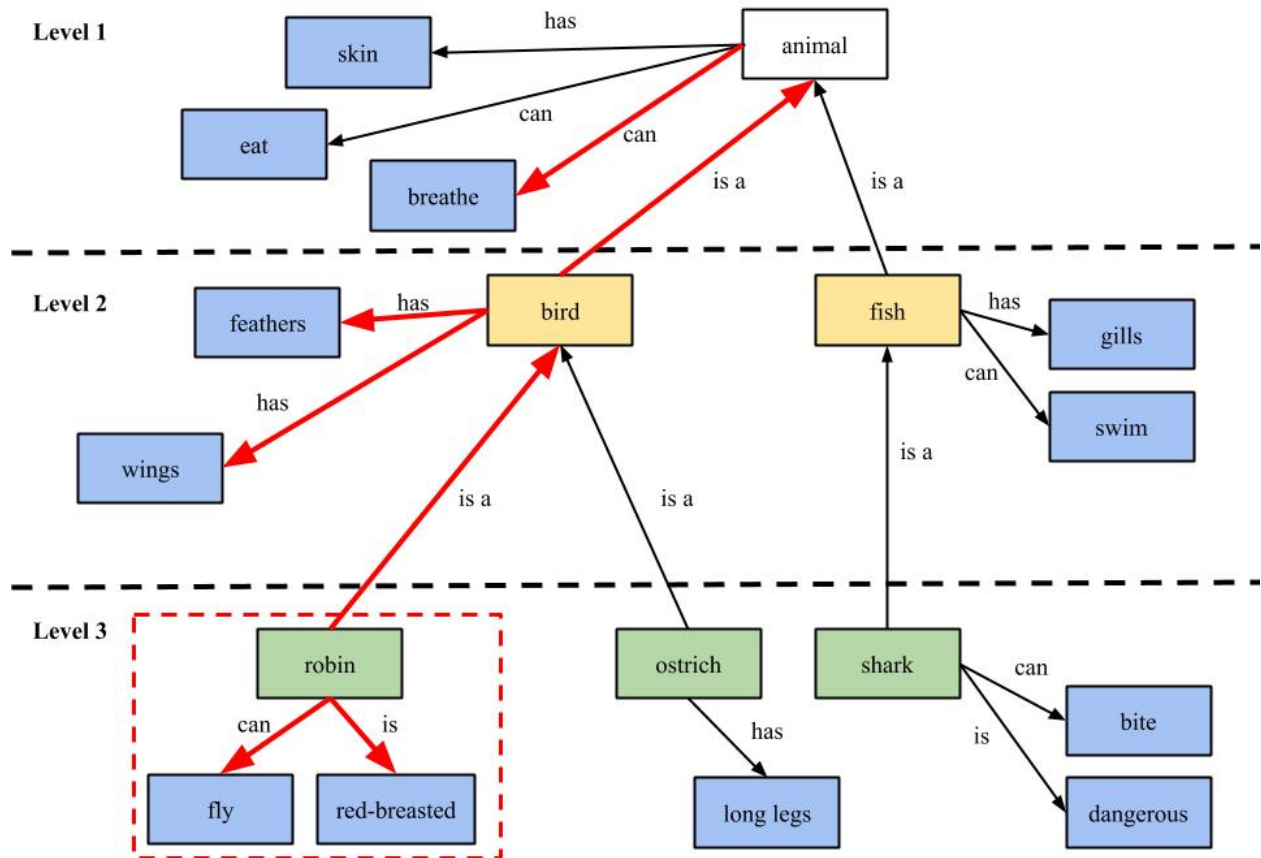


Figure 2.3. An illustration to depict Collins and Quillian's (1969) hierarchical network model. Demonstrating the verification pathways for the sentence "*A robin can breathe*".

This model is appealing in many ways. However, it makes a number of assumptions concerning the hierarchy structure and processing semantic memory that have not held up in experimental tests. The first assumption known as the semantic distance effect, assumes the processing time to transverse the "is a" links increases with the number of inferences made (Meyer, 1970; Rips, Shoben, & Smith, 1973). For instance, propositions about specific properties (e.g., *A robin is a bird*), should be verified faster than propositions about general properties (e.g., *A robin is an animal*). Initial experiments support the hierarchical processing model, basically because such properties were stored higher in the hierarchy and required an increased number of inferences. As depicted in Figure 2.4 it shows that the subject *robin* shares a direct link to its superordinate *bird*, whereas *animal* is across multiple levels of hierarchy, increasing processing time, by the number of inferences required (Collins & Quillian, 1969).

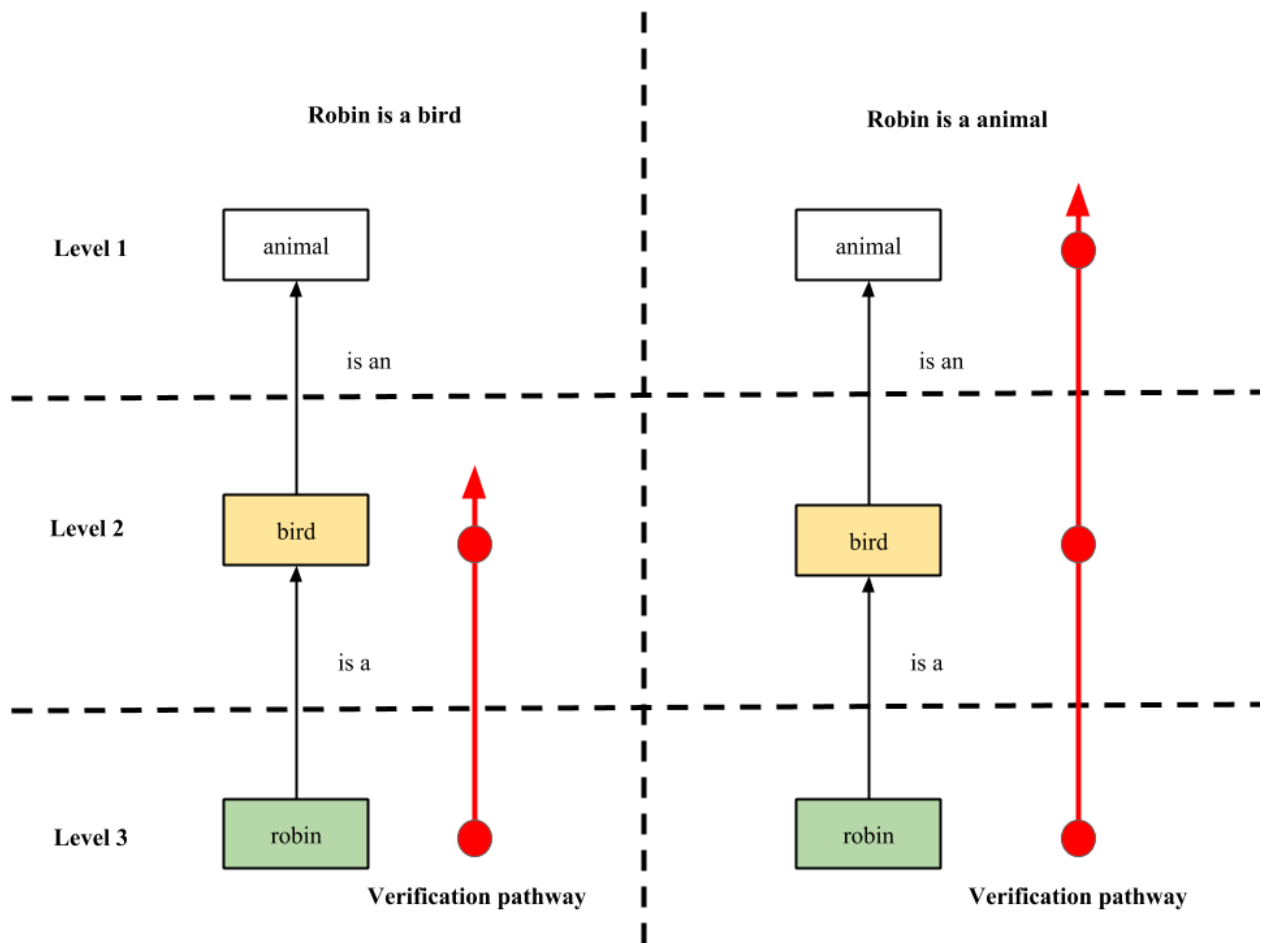


Figure 2.4. An illustration to depict the semantic distance effect between the verification of two different statements.

The second assumption states that statements verifying property attributes of concepts should be verified faster if the subject term is paired with a property stored at a higher semantic level. The theory behind this assumption is also used to explain the idea of cognitive economy storage. This principle assumes that the properties of a concept are not stored at all nodes to which they apply, but at the highest possible semantic level, as these properties are inherited by concepts on a more basic level (Chang, 1986; Conrad, 1972). For instance, *eats* would be stored with *animal*, but not with *bird*, because all animals have to eat, whereas, *fly* would only be stored with *robin* but not with *bird*, because *robins* can *fly* but not all *birds* can *fly*. Therefore, we would expect that the statement “a robin can eat” would be verified faster, when compared to “a robin is red-breasted”, based on the idea of cognitive economy.

The idea of organising conceptual knowledge in a hierarchy model makes intuitive sense. However, issues have arisen regarding disparity in features that all concepts share and features that are typical of a particular concept but are not present in all cases. A study by Smith, Shoben, and

Rips (1974) asked participants to verify the statements “*a robin is a bird*” and “*an ostrich is a bird*”, response times in verification were collected and measured. If the semantic distance effect assumption is valid, verification time should increase proportionately to the number of levels in the hierarchy. Hence, there should be no difference in processing time for verifying both statements within the same level of hierarchy. Consequently, this was not the case as participants demonstrated increased verification time in processing “*an ostrich is a bird*”, even though both animals are birds. A similar study conducted by (Conrad, 1972) exposed issues with the second assumption. Resulting in a verification experiment being conducted to measure the distance between subject term relationships with a property attributes stored at different hierarchy levels. The experiment involved the verification of the statements “*an animal can move*”, “*a fish can move*” and “*a shark can move*”. The assumption is that one would expect “*an animal can move*” to be processed faster, because the property *can move* would be stored closest to the node *animal*. However, the results showed that participants had similar processing response times to all statements, despite the increasing levels of hierarchy from the node *animal* to *fish* to *shark*. Both studies highlight issues with the storage of conceptual knowledge organised in the structure of a hierarchy, that will be addressed in the next section.

#### **2.4 Semantic Network Theory – Spreading activation model**

The second model, proposed by Collins and Loftus (1975), known as the spreading activation model was a revised model that focused less on hierarchical structures and more on the processing of unique concept nodes and their properties. In this theory, the basic notion of spreading activation is raised, meaning that one memory structure may activate another adjacent (related) structure based on a retrieval cue. Retrieval is an *active* process, starting with the target concept and spreading out along the links associated with these concepts. Once activation begins the depth and breadth of the spread occurs in accordance with an individual’s semantic network representation. There is also extensive evidence that suggests successful retrieval is a “memory modifier” having the potential to improve learning by enhancing subsequent encoding based on the concepts activated in the retrieval process (Grimaldi & Karpicke, 2012; Storm, Bjork, & Storm, 2010; Van den Broek et al., 2016). The activation pathways and link distances are paramount in explaining the increased effectiveness of this model. For instance, a concept is more likely to be activated if there is a shorter connection between it and the starting concept or there is a larger number of paths directly or indirectly leading to it from the activation concept (Chang, 1986).



In this model, the researchers have redefined the assumptions and improved the processing and organisation of semantic memory. The first improvement assumes that the retrieval process begins with a node that is most obvious to the retrieval cue. For instance, if students were asked to discuss the concept *vehicles*, a diversity of connections can be activated depending of the experiences of the student. Firstly, is likely that students will recall the most prototypical examples, such as *bus*, *truck* and *car*. Secondly, they may recall concepts that share a relationship to vehicle and are specific examples of the concept, such as *ambulance* and *fire engine*. Finally, more abstract links can be made that are based on episodic experiences, such as *street*, which could relate to the concept *ambulance* based on knowledge of other concepts, such as “a car drives through a street, therefore an ambulance car drives through a street” (Chang, 1986). As shown in Figure 2.5 the more properties that two concepts have in common, the more interconnected the network, aiding in the recall and activation of the knowledge. Network a, provides an example of a richer cluster of related concepts, in comparison to the concept Network b. Consequently, the assumptions that plagued the hierarchical network model, such as cognitive economy and hierarchical organization, have been abandoned in the spreading activation model.

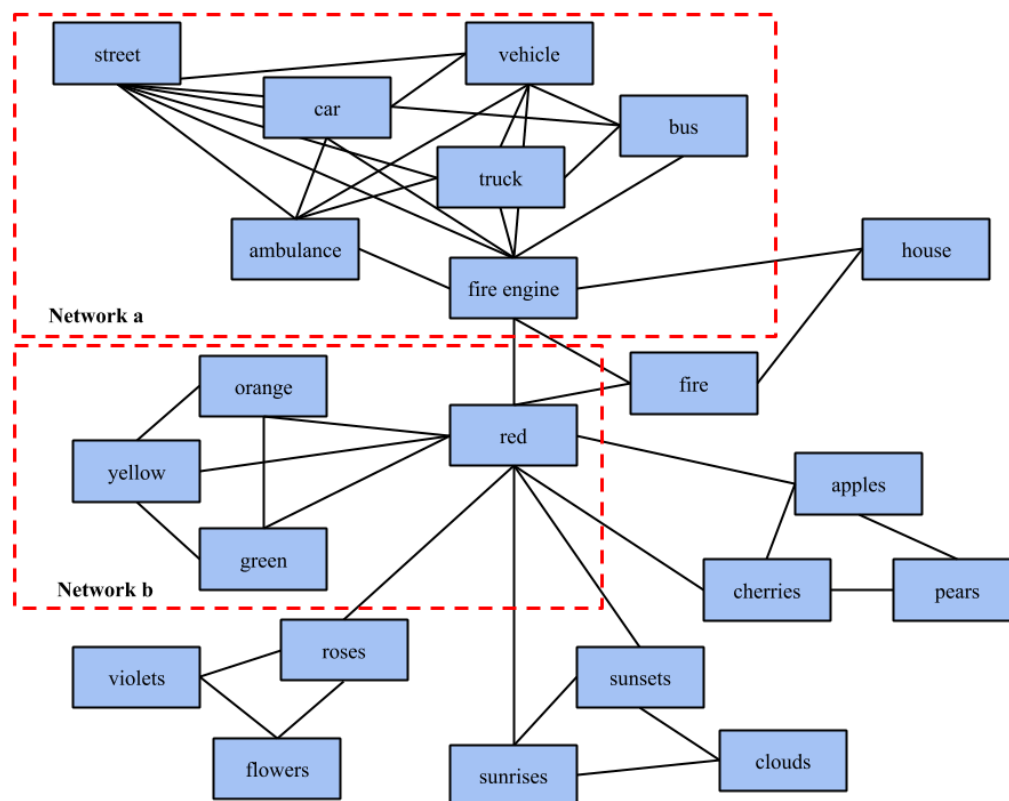


Figure 2.5. A memory representation according to the Collins and Loftus (1975) demonstrating spreading-activation theory. This illustration shows different clusters of concepts across two internal networks.

The second improvement established by Collins and Loftus (1975) is that semantic memory is organised along the lines of semantic similarity. The key premise of semantic similarity is that the more properties that two concepts have in common, the more closely related and stronger their relationship in the semantic network. This may provide a justification for the results obtained in the Smith, Shoben and Rips (1974) study as the concept *robin* shared more common properties with *bird*, than the concept *ostrich* as shown in Figure 2.6. This illustration shows that activating the concept *robin* will activate three attribute features of bird, whereas, the concept *ostrich* will only activate two attribute features. Furthermore, the memory representation shown in Figure 2.8 helps understand this improvement. For example, the nodes for different vehicles (*fire engine*, *ambulance*, *truck*, etc.) are identified as having a strong connection, considering the common properties that they share such as *has at least four wheels*, and they can be used to transport people from place to place and more. Conversely, nodes like *fire engine*, *apples*, *blood*, and *roses* are not as similar because they only have one property in common, which are being the colour red. The more similar the two nodes are based on their common properties, the closer in proximity they will be represented in the semantic network like, for instance the distance between the nodes *fire engine* and *truck* is much closer compared to the distance between the nodes *fire engine* and *cherries*.

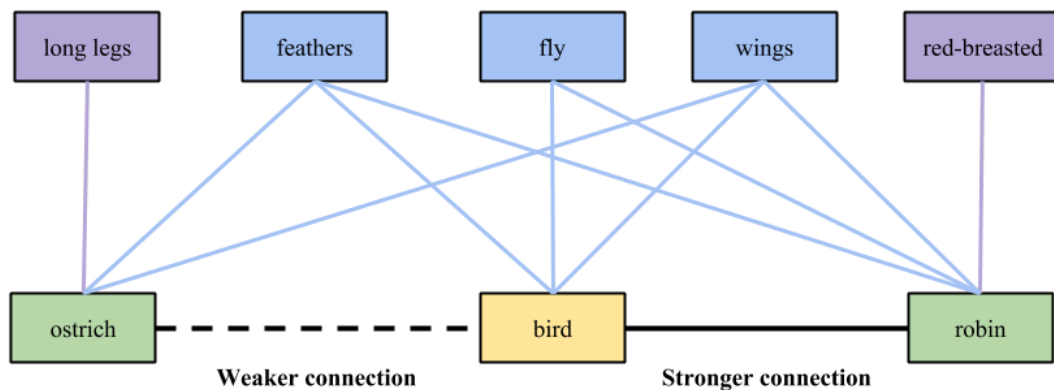


Figure 2.6. An illustration to help explain the typical differences of similar concepts.

The most significant difference between this model and the hierarchical network model lies in the processing of concepts. Collins and Loftus (1975) propose a search process that traces all links simultaneously by means of spreading activation, rather than moving from one level of the hierarchy to the next. Triggered by an input word (or concept), the spreading activation then expands constantly, first to all the nodes linked to the first node, then to all the nodes linked to each of these nodes, and so on. To demonstrate this, Figure 2.7 shows a semantic network of

concepts based on a year ten secondary school science topic the “periodic table”. Activation would begin with input node (i.e., *the periodic table*) and expand to its immediate predecessor nodes like *elements*, *groups*, *periods* and *symbol*. These concepts identified are the key concepts in learning the periodic table. Activation continues to spread further from these nodes (e.g., *elements*) to other nodes like *atoms* and *examples*, until all concepts in the network are activated. In summary, this semantic network serves the purpose of a visual illustration representing all the ideas that students are expected to learn out of the topic.

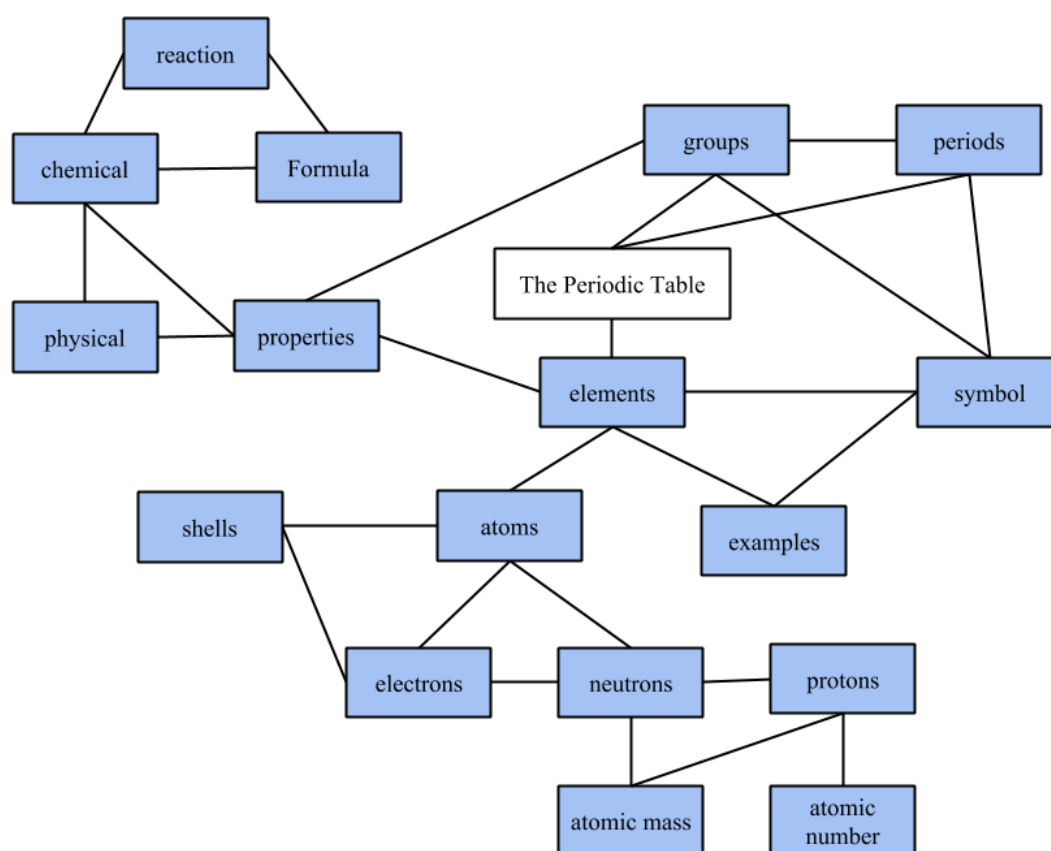


Figure 2.7. A schematic representation of the concept *Periodic table* based on a year ten secondary school science topic.

This same theory is used to support the recall mechanisms in the Concept Retrieval Technique. That is, when one is asked to recall the relevant concepts of a topic in question, the activation of concepts will spread from the first given concept (i.e., the topic itself) to the concepts that are directly connected to it, and from each of these concepts to subsequent others. However, two main assumptions to the processing of concepts in this theory can be made. First, when a

concept node is being processed, activation spreads out along the links in the network in a decreasing gradient. This decrease is inversely proportional to the strength of the links in the paths of the network. One can think of the activation like a signal from a source that weakens as it travels outwards. The second assumption about the processing of concept nodes in this theory is that only one concept can be processed at the start. This means that activation can only start at one node at any point in time but will continue simultaneously to the other nodes that are encountered as it spreads out from the node of origin. For instance, referring to the example in Figure 2.7 there can only be one node of origin (the periodic table). Then, starting from this node, the spread of activation will expand simultaneously to the other nodes. Subsequent models have been developed to address some of the assumptions identified with the spreading activation model.

These include, the emergence of feature models, which propose that for each concept a feature list can be generated. Features are fundamental in classifying objects, forming concepts, and making generalizations (Tversky, 1977). Furthermore, concepts are linked by their features and can be categorised by means of their similarity on critical features (Rosch & Mervis, 1975; Smith et al., 1974). For example, the concept *shark* could be represented using important features, such as *can bite*, *is dangerous*, *has gills* and *can swim*. Whereas, an ostrich would be represented by a different set of features, such as *has feathers*, *has wings*, and *has legs*. This model assumes that words or their conceptual counterparts exist as independent units in semantic memory, connected in a network by labelled relations. Processing requires retrieval of the stored relations between concepts and a comparison of these relations to that asserted in the propositions.

Following on from feature models, associative models were proposed that also assume all concepts are linked, but the linkage is established and enhanced if two concepts are simultaneously active in memory (Raaijmakers & Shiffrin, 1981). When simultaneously activated, their association is stronger than with other concepts that were not simultaneously activated. During the recall process the brain reactivates the original neural representations, generated in the encoding process, the strength of these representations determines how quickly the knowledge can be recalled. This theory symbolises long-term memory as a complex network of connected nodes and memory formation is a continuous process in which neural connections are stabilized over different learning episodes (Atkinson & Shiffrin, 1968). Finally, recent advances in technology have enabled connectionist models to be developed that rely on the same foundations of semantic knowledge built from the previous two models (i.e., how it is represented, organized, and processed), but help address issues raised concerning the complexity of semantic networks.

## 2.5 Semantic Network Theory – Connectionist model

Contrary to early models of semantic networks that describe a single association, modern connectionism models describe large multilevel representations, referred to as parallel distributed networks, that help describe how knowledge representations might interact with other cognitive processes. In general, connectionist models have been constructed to explicitly capture the dynamic nature of semantic knowledge, through the analysis of neuron activation patterns by performing cognitive tasks (Joanisse & McClelland, 2015; Jones et al., 2015). These results reveal that knowledge representations are distributed across the entire neural network, meaning no single neuron uniquely encodes a concept or category (Rogers & McClelland, 2004). Furthermore, these distributed representations assume that concepts and their relations are patterns of activation that represent knowledge in terms of weighted connections between interconnected units within the brain (Rogers & McClelland, 2004). Finally, semantic knowledge is acquired through the gradual changes in the strength of these connections within the interconnected units, in response to processing external inputs from the environment (Joanisse & McClelland, 2015; Rogers & McClelland, 2004).

Rumelhart and Todd (1993) proposed the *feed-forward* connectionist model to demonstrate that the hierarchical propositional network structure addressed in previous models could also be captured in distributed representations. The typical layout of this connectionist model is that it consists of three layers: an input layer, a hidden layer, and an output layer. Within the input and output layers are individual nodes that correspond to the individual components of each proposition. The structure of a simple proposition is reflected in the architecture of the network by the *item-relation-attribute* form. For instance, a proposition is represented by a concept within the first (item) slot, relation or linking terms then occupy the second (relation) slot, and the attribute values occupy the third slot. Each item is represented by an individual input unit in the layer labelled *item*, different relations are represented by individual units in the layer labelled *relation*, and the different possible completions of the three element propositions are represented by individual units in the layer labelled *attribute* (Rogers & McClelland, 2004, 2008). When the model receives an *item* and *relation* pair in the input layer, the network's job is to activate valid completions of the proposition within the *attribute* units in the output layer. For example, Figure 2.8 illustrates the input unit's *robin* and *can* being activated in the input layer. As a result, each stored proposition is represented in the network by a unique pattern of activity across each unit, depending on the strength of connection, networks must learn which attribute units need to be activated based on the input layer (Rogers & McClelland, 2004, 2008)

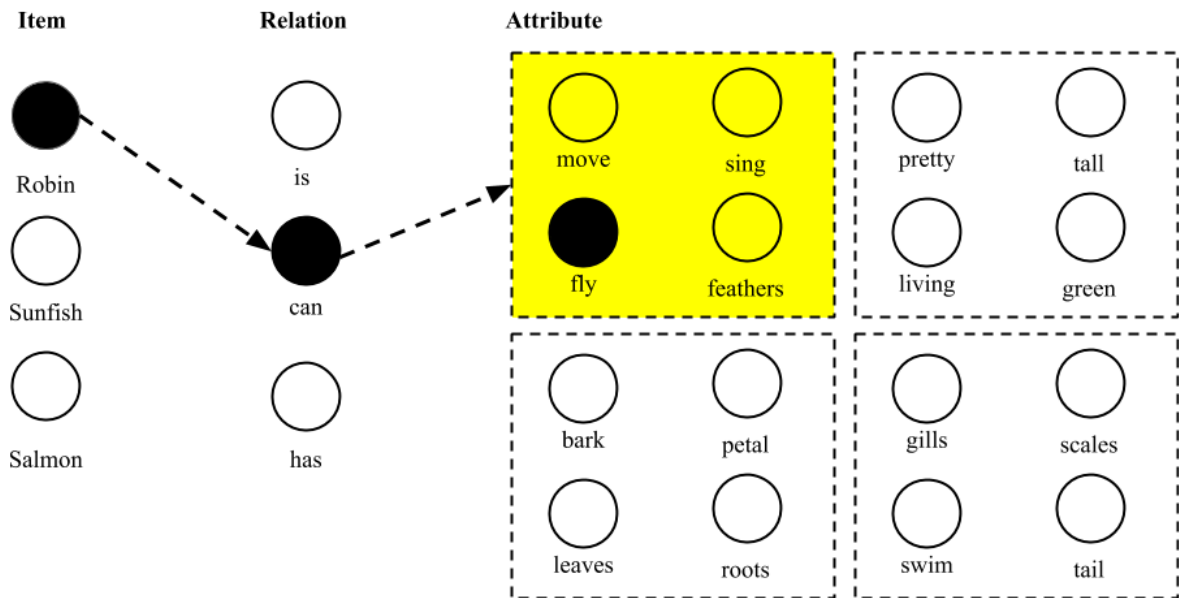


Figure 2.8. An illustration showing the item-relation-attribute connection for the statement “a robin can fly”.

Most connectionist models also have one or more sets of intervening units between the input and output units, which are often referred to as hidden layers (Jones et al., 2015). The hidden layer is the critical property that distinguishes this model from earlier models of semantic memory and they create internal representations for all possible inputs. In Figure 2.9 the unique connections between the input, hidden and output layers are depicted. The network consists of a series of nonlinear processing units, organized into layers, and connected in a feed-forward manner. The input layer is activated first, through the interaction with objects in the world (i.e., *seeing a robin fly*) or of spoken statements about these objects (i.e., *can a robin fly?*). Patterns are generated by activating one unit in each of the *item* (e.g., *robin*) and *relation* (e.g., *can*) layers. It is assumed that the signal is distributed throughout the network through some form of spreading activation. It then passes through the hidden layer, modulated by connection weights that are adjusted, before spreading to the outer layer to activate an appropriate attribute representation. The network learns to associate these two sets of inputs (i.e., *robin + can*) with an output representing semantic features (*fly, move, grow, etc.*). For example, *robin, oak, salmon, and daisy* all use the same hidden units. However, what differentiates their internal representations is that they instantiate different distributed patterns of activation within the hidden layer (Jones et al., 2015). The assumptions concerning the existence of spreading activation and the complexity of weighted connection strength within the hidden layer needs further discussion.

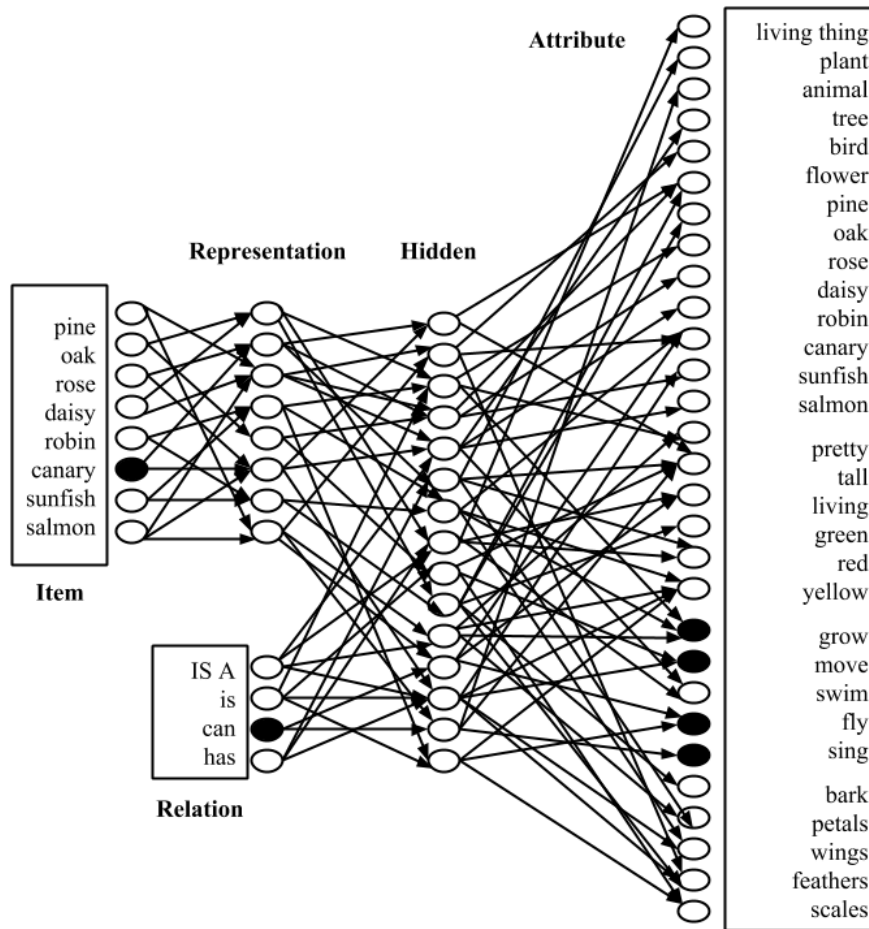


Figure 2.9. A connectionist model of semantic memory adapted from Rumelhart and Todd (1993).

Similar to the spreading activation theory proposed by Collins and Loftus (1975) connectionist models consist of layers of immensely interconnected units and involves the idea of spreading activation. Hence, the activation of a single item is fed into the input units, and that activation in turn activates (or suppresses) other connected units, as a function of the weighted connection strength between each unit to produce an output (Jones et al., 2015). Specifically, the weights connecting the *item* and *representation* units evolve during learning, so the pattern of activity generated across the *representation* units for a given item is a learned internal representation of the item. Furthermore, the hidden layer feature is unique to connectionist models, which allows the system to explain more complex cognitive tasks. In fact, it is so flexible that it can be seen as a unifying theory as it assumes that all types of mental knowledge can be understood within it, the concepts, how they are connected, how they are processed (through spreading activation), and how they represent mental processes.

In summary, all models assume that semantic networks exist and are based on a connection between concepts. The Concept Retrieval Technique is based on the idea that a person can recall multiple concepts after another based on the strength of linkages that exist between the recalled concepts (Collins & Quillian, 1969; Shavelson, 1972). The hierarchical model provided the foundation to help visualise how conceptual knowledge could be represented in a connected network form. Specifically, the assumption that semantic knowledge is stored representations of concepts, linked together in a taxonomically organized hierarchy. The spreading activation model developed from contradicting evidence that suggest that the processing time of concepts does not occur according to such a hierarchy. This model deemphasized the hierarchical nature of the network model in favour of the process of spreading activation through all network links simultaneously accounting for the semantic priming phenomena (Jones et al., 2015). The spreading activation model provided concrete evidence that concepts do not exist in isolation but they exist in the forms of networks and processing of these concepts occurs as a form of spreading activation throughout these networks. While the hierarchical network model and the spreading activation model are one of the older theories of network representations, more recent models (e.g., connectionist models) have been developed to explain more complex phenomena like language acquisition, neural activity of semantic memory. In addition, to distinguishing the underlying principles of different semantic memory models, neuropsychological evidence can be examined concerning how effective retrieval and activation patterns in the brain support the Concept Retrieval Technique. In the next section, the literature review will examine empirical research reinforcing how semantic knowledge is represented, organized and processed in the brain.

## **2.6 The neuropsychological evidence of Semantic Network Theory**

The human brains acquire and use concepts with such apparent ease that neuroscience is often taken for granted. The domain of semantic memory consists of stored representations about features and attributes of concepts, and the processes that allow the effective retrieval of this information in language and thought (Martin & Chao, 2001). This critically important system of brain is involved in a wide range of cognitive functions including assigning meaningful representations to words and sentences, recognising objects, information recall from learnt concepts and new information acquired from experiences. One of the basic goals of cognitive neuroscience is understanding how the brain represents and processes semantic knowledge, the regions of the brain involved in the retrieval and the activation of semantic knowledge from neural



connections and networks. The development of a more concrete understanding of semantic memory began with a period known as the “cognitive revolution”.

Pioneering these developments was Hebb (1949) who studied the impulses in the brain, proposing that networks of concept representations existed in long-term memory. These representations are organized through shared elements of larger cell assemblies, that are not localised to one area of the brain (Baddeley, 1992; Campoy, Castellà, Provencio, Hitch, & Baddeley, 2015; Eichenbaum, 2017; Schunk, 2008). Miller (1956) quantified the capacity of short-term memory to be seven items (plus or minus two), emphasizing the importance of encoding and recoding associated “chunks” in long-term memory (Chen & Cowan, 2005). The notion of “chunking” is described as a collection of items or network that depending on the retrieval cue, may be cycled within short-term memory activating new or retrieving existing semantic network representations (Cowan, 2010; Shiffrin & Nosofsky, 1994; Unsworth & Engle, 2007). Specifically, the hippocampus is responsible for processing information chunks from short-term memory, storing them in long-term memory and maintaining associative connections with other semantic networks to avoid memory decay (Schunk, 2008; Wolfe, 2001)

Tulving (1972) promoted a distinction between episodic and semantic memory, with episodic memory being unique, concrete and personally experienced events that occur in a unique spatial and temporal context (Tulving, 1986, 1993). Whereas, semantic memory is an individual’s store of knowledge such as facts, meanings and concepts, abstracted from many experiences and not dependent on any specific event (Binder & Desai, 2011; Buckner, Wheeler, & Sheridan, 2001; Eichenbaum, 2017). This research helps justify that the better integrated and the more comprehensive these semantic networks are, the more concepts can be recalled. Put simply, a student with relatively more knowledge and tighter links among concepts will be able to freely recall more concepts than a student who possesses a less extensive semantic network with weaker links between concepts.

An old idea in behavioural neurology has been that concepts are defined by sensory and motor attributes and features acquired during experience. However, it has been suggested that concepts may be represented in the brain as a distributed network of sensory, motor and more abstract functional information. Prior to the advent of *functional magnetic resonance imaging* (fMRI), our knowledge of the neural bases of semantic memory was dependant on studies of patients with brain injury or disease. Patients who displayed damage to the temporal lobes have demonstrated difficulty in naming objects and retrieving information about object-specific characteristics. This suggests that object-specific information may be stored, at least in part, in the temporal lobes (Martin & Chao, 2001). The examination of the Concept Retrieval Technique as

an authentic assessment measure is dependent on neuroscience research, that provides a clear understanding of how the brain physically accesses cognitive processes in the recall of memory representations (Grimaldi & Karpicke, 2012; Roediger & Guynn, 1996; Tennyson & Rasch, 1988). For example, *functional magnetic resonance imaging* (fMRI) has been used to capture brain activation patterns based on different memory tasks (Campoy et al., 2015; Wolfe, 2001). Research utilising this technology has demonstrated correlations between hippocampal and prefrontal activity, with the hippocampus identified as the hub of brain activity that supports encoding and recall of memory representations (Buckner et al., 2001; Campoy et al., 2015; Eichenbaum, 2017; Shuell, 1986; Sternberg, 1984; Wagner, 1998).

The core argument presented in recent neuroscience research is that semantic memory consists of both modality-specific sensory representations and the existence supramodal representations that support a variety of conceptual functions including object recognition, social cognition, language and the uniquely human capacity to construct mental simulations of the past and future (Binder & Desai, 2011; Thompson-Schill, 2003). For example, our knowledge of a *dog* would include its attributes such as its visual features, sounds such as its bark, and even its texture and smell. Each attribute activates a different part of the brain, such as visual features will be represented in the regions involved in the processing of visual forms, its bark will correspond to the auditory regions, texture and smell will involve the tactile representations. This recognises that the neural representation of how an object looks, how it moves, how its texture feels like and so on contributes to the idea that conceptual knowledge is indeed a widely distributed neural network (Patterson, Nestor, & Rogers, 2007). A meta-analysis of 120 functional neuroimaging studies, focusing on activation sites in the brain, during the cognitive act of accessing stored semantic knowledge, discovered consistent application across up to seven regions (Binder & Desai, 2011).

Apart from neuropsychological evidence that semantic representations are stored in the anterior temporal lobes that acts as a hub linking to the knowledge of the other attributes like visual and perception, there is also empirical evidence that demonstrates the semantic processing of spreading activation in human's semantic memory (Martin, 2007). According to Neely, Keefe, and Ross (1989), semantic priming is a process that contributes to the automatic spreading activation among related words. When a prime word is processed, its corresponding node is activated and this activation then spreads to related nodes in the network and this results in a shorter response time for related targets due to reduced efforts for recognition (Collins & Loftus, 1975). In semantic priming, a word stimulus is used to initiate the activation of a network and its neighbouring semantic associates. Each word pair shares semantic attributes but are different in orthographic

and phonological aspects. The results from a neuroimaging study by Wible et al. (2006) corresponded to this prediction. The study manipulated the levels of semantic priming by using different connectivity word pairs. It was hypothesized that pairs with high connectivity would have the highest levels of semantic priming and no semantic priming was expected for unrelated word pairs. Figure 2.10 shows the difference between word pairs with high and low levels of connectivity. This study involved quantitative calculations to determine the level of connectivity. For instance, the calculation for the concept *dinner* involved computing the number of connections (17) and dividing the number by the number of associates (5) resulting in a connectivity score of 3.4.

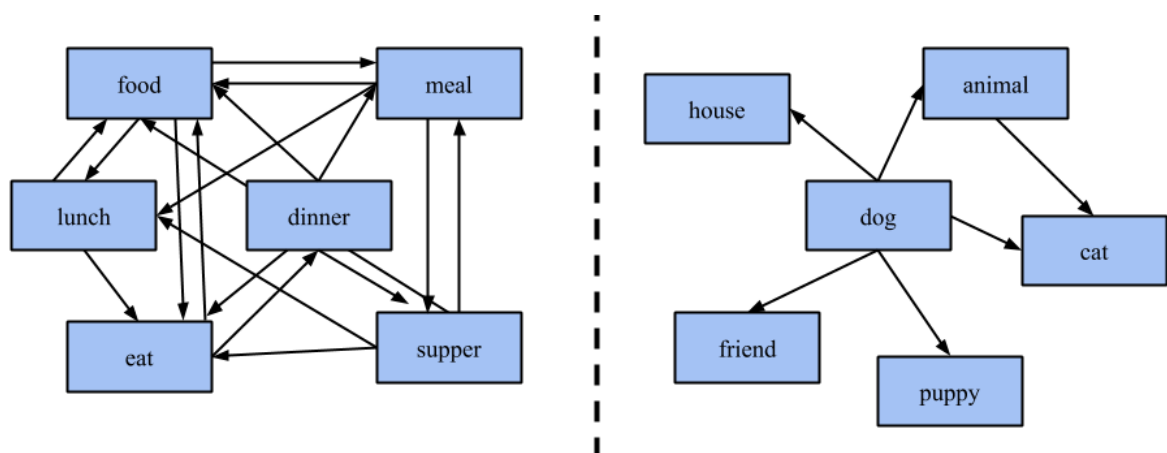


Figure 2.10. An illustration showing the connectivity between associates. The concept *dinner* shows a high degree of connectivity between associates and the concept *dog* shows a low connectivity.

Indeed, reaction time was the fastest for word pairs of high connectivity and the slowest for unrelated word pairs. In addition, fMRI analyses showed significant differences as a function of the connectivity among word pairs. That is, activity was found to systematically decline as the semantic priming of word pairs increased with connectivity. This result was consistent to many other research studies that have reported the activity of brain regions related with semantic priming (Copland, de Zubizaray, McMahon, & Eastburn, 2007; Rossell, Price, & Nobre, 2003; Sachs et al., 2008).

In summary, there is empirical evidence in the neuropsychological studies presented that make evident, how semantic memory is represented and processed in the human mind. The results suggest that the anterior temporal lobe plays a central role in the semantic representations in the brain and that the processing of semantic memory occurs through the form of automatic spreading activation. As long as semantic memory is intact, the automatic semantic processing in humans

should not be impaired. Despite the significant influence that semantic network theory has had on the field of psychology and cognitive neuroscience, there is limited research about its application in the field of education. In the next section, the literature review will address the relevance of semantic network theory in educational assessment and testing settings.

## **2.7 The Concept Retrieval Technique and its application to education**

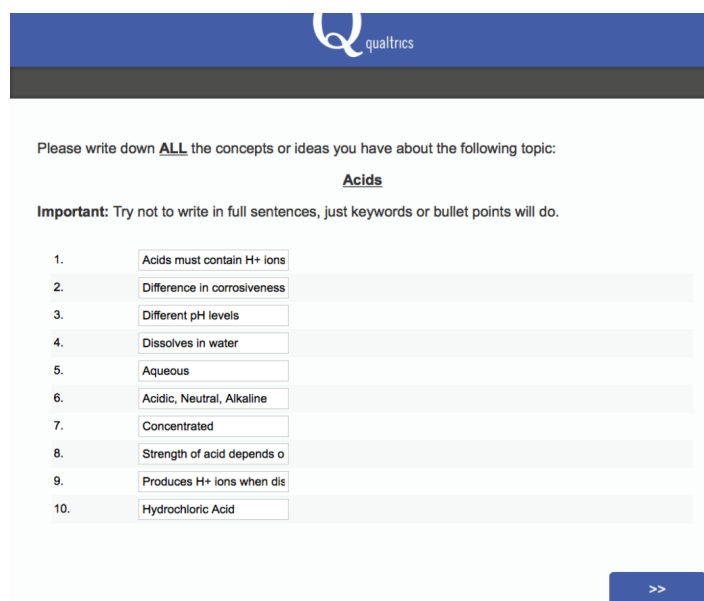
Although the theory of representing conceptual knowledge in semantic networks has been a major influence especially in the field of psychology and cognitive neuroscience, there is limited research about its application in the field of assessment. There have been several attempts in the past two decades to achieve this objective. The most apparent approach of applying the theory of semantic networks in education is the use of concept maps (Novak, 1990). Concept maps, introduced by Novak and Gowin (1984), are propositional diagrams that include concepts and their relationships. The objective of concept maps is to assess the depth and breadth of a student's semantic knowledge structures. There are a variety of ways to administer a concept map, primarily they instruct the test-taker to recall all relevant concepts pertaining to a topic in question and to draw property or causal relationships between them (Novak, 1990; Novak & Gowin, 1984). Although concept maps are appropriate with regard to measuring retrieval rather than recognition, educators and psychometricians have raised concerns regarding their scoring reliability (Eppler, 2006; McClure, Sonak, & Suen, 1999; West, Pomeroy, Park, Gerstenberger, & Sandoval, 2000).

There are a number of different ways to administer concept maps with most commonly used techniques being the “construct-a-map” and the “fill-in-the-blanks”, especially within education (Ruiz-Primo, Schultz, Li, & Shavelson, 2001). A study comparing these two mapping techniques found that the low-directed “construct-a-map” technique imposed a high cognitive demand on the retrieval of conceptual knowledge from students. Unfortunately, this technique is considered problematic as students need adequate training in concept map construction. Furthermore, scoring is difficult, time consuming and reliability concerns have been raised (Ruiz-Primo et al., 2001). On the other hand, the high-directed “fill-in-the-map” technique, is a simpler alternative in application and scoring procedures. However, by providing an initial representation to students imposes a low cognitive demand as students are only required to demonstrate recognition and guessing can also occur (Ruiz-Primo et al., 2001).

Besides the issues with training students in the construction of concept maps, the more serious concern resides in its scoring procedures. This is because each concept map is individually constructed by the test-taker representing the person's cognitive organization, which causes it to

be highly idiosyncratic with hundreds of possible concept map permutation (Ho, Kumar, & Velan, 2014; Ruiz-Primo & Shavelson, 1996). To reliably score concept maps, a marking scheme needs to be devised capturing all possible relationships. Furthermore, scoring procedures may emphasize the organization of concepts (Novak & Gowin, 1984), interconnectedness of concepts (McClure, Sonak, & Suen, 1999), or the validity of the propositions used (Nicoll, Francisco, & Nakhleh, 2001). Consequently, the valid and reliable use of concept maps as an assessment tool is found to be unattainable and currently concept maps are predominately used as teaching tools to help students identify gaps in their understanding and to provide formative feedback, rather than to be used for assessment purposes per se (Daley & Torre, 2010; Edmondson, 2005).

Although the Concept Retrieval Technique is similar to concept maps in its intention to measure a person's conceptual knowledge, it differs significantly in how it is operationalized. Instead of drawing a complete map, the Concept Retrieval Technique requires the test-taker only to write down all the concepts and ideas that come to mind regarding a given topic or subject. Moreover, as shown in the example in Figure 2.11 the test-taker is instructed to only write down keywords or bullet points (i.e., concepts without propositions) and not full sentences to make scoring more straightforward and reliable. The mental elaboration during the attempt to retrieve a target from memory extends the semantic network of the tested information by creating or strengthening connections with related concepts, they also activate several semantically related words while searching for the target, this semantic elaboration during retrieval provides additional retrieval routes (Carrier & Pashler, 1992; Pyc & Rawson, 2010; Van den Broek et al., 2016).



The screenshot shows a web-based interface for the Concept Retrieval Technique. At the top, there is a blue header with the 'qualtrics' logo. Below the header, the instructions read: 'Please write down ALL the concepts or ideas you have about the following topic: Acids'. An important note states: 'Important: Try not to write in full sentences, just keywords or bullet points will do.' Below this, there is a list of 10 numbered items, each with a text input field containing a student's response. At the bottom right, there is a blue button with the text '>>'.

1.	Acids must contain H <sup>+</sup> ions
2.	Difference in corrosiveness
3.	Different pH levels
4.	Dissolves in water
5.	Aqueous
6.	Acidic, Neutral, Alkaline
7.	Concentrated
8.	Strength of acid depends o
9.	Produces H <sup>+</sup> ions when dis
10.	Hydrochloric Acid

Figure 2.11. An example of the responses of one student to and online Concept Retrieval Technique administration for the topic “acids”.

This flow of recalled concepts in the network is in line with the scientific theory of “spreading activation”, in which the activation of one concept gradually spreads through the links to other related concepts in the network (Anderson, 1983; Collins & Loftus, 1975). The more concepts a person has about the topic, the wider the spread of activation. The Concept Retrieval Technique is hypothesized to work on the same principle – to recall concepts of a topic; activation of the first concept (i.e., the topic) will spread to adjacent concepts until all concepts in the knowledge network are activated. Although the spread of activation can theoretically go on indefinitely, in practice, the Concept Retrieval Technique only measures a limited number of concepts that a student perceives as most relevant to the topic in question.

For this reason, the target word list developed by the subject matter experts plays an important role in determining which concepts are relevant for understanding the topic. Following from this, a student with extensive knowledge about a topic will be able to activate more relevant concepts than one who has little to no knowledge about the same topic. This is because the former would have more concepts accumulated which means that there are more available concepts in this student’s knowledge network—hence, the increased number of relationships among relationships would determine a wider spread of activation. In addition, this spread of activation will continue for as long as the student perceives the concept to be relevant to the topic. In short, the scores of the Concept Retrieval Technique are hypothesized to represent the knowledge network of an individual, given that students with extensive knowledge about the topic would have more accumulated concepts and thus a wider spread of activation in the semantic network.

In summary, the Concept Retrieval Technique is underpinned by established and contemporary models from cognitive psychology and neuropsychological evidence, supporting the fundamental idea that conceptual knowledge is encoded and retrieved from semantic memory networks. That is, students’ conceptual knowledge for a topic in question is represented as concepts and organized this network structure. When students get the chance to learn about a topic in question and acquire new concepts, their network of concepts becomes more extensive. Then, when the student is asked to list down relevant concepts in the Concept Retrieval Technique, a given concept (i.e., the topic in question) prompts the activation of concepts that are connected immediately to this given concept. Subsequently, from each of these concepts, activation will continue to spread to the other connected concepts until it reaches a point to which the student perceives the concept as not relevant to the topic anymore. This is the general premise of the Concept Retrieval Technique that students with more knowledge will naturally be able to write down more relevant concepts than a student with less knowledge. Furthermore, the Concept

Retrieval Technique has the potential to address the validity and reliability shortcomings of other conceptual knowledge assessments, such as concept mapping and educational assessment in general. The terms validity and reliability are important in the field of assessment.

The development of the psychometrics theory that involves validity and reliability began in the early 20<sup>th</sup> century for the purposes of large-scale testing in the field of social sciences but was later commonly applied to classroom assessment. Because classroom assessment provides substantial information to the students, teachers, and parents, it is critical that classroom assessments are adequately developed so that accurate inferences can be made from its results. The next chapter will elaborate on how to validate the Concept Retrieval Technique as a measure of students' conceptual knowledge.

The background of the entire slide is a complex, abstract network diagram. It consists of numerous circular nodes, some of which are white and others are dark gray. These nodes are interconnected by a web of thin, light gray lines, creating a sense of connectivity and structure. The overall aesthetic is modern and technical, typical of a presentation on computer science or data management.

# 3

**Chapter**

## **The Reliability of the Concept Retrieval Technique**



### Abstract

The objective of the present studies was to establish reliability of a new approach to measuring conceptual knowledge, the Concept Retrieval Technique. This test requires test takers to freely retrieve from memory concepts they feel are relevant to a given topic, with raters awarding one mark for each correctly retrieved concept when matched against a target word list. Four studies were conducted. In Study 1 ( $N = 73$ ), inter-rater agreement for the marking procedure of the CRT was determined by means of Cohen's kappa. The kappa was  $\kappa = .85$ . Study 2 ( $N = 211$ ) explored how consistent inter-rater agreement was across different subjects, age groups, and raters. Mean kappa was  $\kappa = .92$  suggesting that the Concept Retrieval Technique can reliably be marked. Study 3 ( $N = 62$ ) provided evidence concerning whether there were any differences between having to read and interpret the full sentences or identifying only the correct concepts in the responses. The results of the studies suggest that the Concept Retrieval Technique is a reliable instrument to measure conceptual knowledge.

### 3.1 Reliability evidence for the Concept Retrieval Technique

Reliability refers to the extent to which an experiment or any measuring procedure yields the same results on repeated trials (Carmines & Zeller, 1979). In Study 1, the reliability of the Concept Retrieval Technique was assessed to determine how reliable the Concept Retrieval Technique is in assessing students' conceptual knowledge. To establish reliability evidence for the Concept Retrieval Technique, recommendations by Messick (1995) and Borsboom, Mellenbergh, and van Heerden (2004) were followed. For a test to be considered reliable, it must be consistent and accurate across testing conditions, including time, samples and ages. The most meaningful indicator of reliability for free recall-type of questions, is the extent to which independent raters agree on assigning marks to the responses (Leiva, Ríos, & Martínez, 2006). There are many different types of reliability evidence but interrater reliability is considered the most crucial for the Concept Retrieval Technique. For tests that require raters to provide some form of judgments, ratings or scores for a specific behaviour or performance, the consistency among raters constitutes an important source of measurement precision. It is to be noted that only inter-rater agreement is of significance because the Concept Retrieval Technique resembles the nature of open-ended questions such that it involves retrieval rather than recognition (as in conventional multiple-choice or true/false questions) when it comes to completing the test.

There are many different approaches to measure interrater reliability. The most straightforward analysis is the percentage agreement of both raters providing the same scores.

However, the analysis of percentage agreement only considers the values that are observed and is unable to reflect the true scores provided by raters. In other words, there is a percentage of scores that could be identified as reliable due to chance and this not considered when percentage agreement is calculated. Cohen's kappa was then introduced to overcome the abovementioned shortcoming. Symbolized by the Greek symbol,  $\kappa$ , this statistic takes into consideration the amount of agreement by raters that is expected to have occurred by chance. The calculation is based on the difference between how much agreement is actually present in comparison to how much agreement would be expected to be present by chance alone.

In Study 2, it was assessed how stable the reliability measures were across Concept Retrieval Techniques administered for different school subjects and age groups. Generalizability validity examines the extent to which a test can be generalized across different groups, settings and tasks (Campbell & Stanley, 2015; Christensen, 2004). The evidence of the generalizability of a test is established through the degree of correlation of the test with a similar construct or aspects of the construct across different tasks, occasions, and raters (Brennan, 1992). As a new measure, it is important that the reliability of the test is measured across different conditions to ensure there is no perceived bias that is affecting the overall result (i.e., Primary vs. High school students or Science vs. Geography). Therefore, to achieve reliability for such questions, the most important part lies in how consistent the raters are in providing the same scores even across different conditions (i.e., different topics and different age groups).

In Study 3, because the scoring approach designed for the Concept Retrieval Technique required raters to look out for correctly recalled concepts, it was unclear if the same scoring approach could be used to score responses in full sentences. Especially, considering that even when students were given clear instructions to write down only in keywords or in bullet points, many provided full sentences. For this reason, this study was designed to determine if the scoring method of identifying correct concepts in test-taker responses, based on a predetermined target word list was significantly different from the scoring method of reading and interpreting the full responses in the Concept Retrieval Technique. To generate an effective reliability analysis this study was presented as an "independent study", increasing the readability and clearly presenting each aspect of reliability evidence.

### 3.2 Study 1: Inter-rater agreement while scoring the Concept Retrieval Technique

In Study 1, we conducted a Concept Retrieval Technique assessing the knowledge acquired by students at the conclusion of the topic “acids”. Two independent raters were used to generate a total score for each student and the results were analysed to determine the consistency of the inter-rater agreement among markers. This was done by generating Cohen’s kappa ( $\kappa$ ). If the kappa showed almost perfect agreement, it would support its overall utility as a reliable classroom assessment tool.

## Method

### Participants

Seventy-three secondary school science students participated (43% female) from three governmental schools in Singapore. Their average age was 14 years ( $SD = .44$ ). Participation was voluntary and they did not receive any compensation for their contribution.

### Materials

A Concept Retrieval Technique was developed for the topic of “acids.” A subject-matter expert worked together with the researchers to generate the target word list containing all admissible concepts for this topic. Furthermore, a review of the learning objectives and materials to be studied ensured that they addressed the knowledge assessed at the culmination of the lesson. See Table 3.1, for an overview of the target word list for this subject. Finally, students were asked to recall ten concepts relating to the trigger concept “acids” as shown in Figure 3.1.

Table 3.1

*The target word list of concepts used for the topic properties of acids.*

Admissible concepts	
corrosive	hydrogen ion
strength	acidity
concentration	dissociation
dissolve	

Figure 3.1. When students do not fill in 10 concepts for the Concept Retrieval Technique, they will not be able to proceed to the next page.

## Procedure

Students participated in a 30-minute lesson on the topic of “acids” within the science subject domain of chemistry. The lesson consisted of a short lecture by the teacher, followed by a self-study session in class with written materials also being provided to students. After the lesson, the Concept Retrieval Technique was administered using the online survey platform Qualtrics® (Provo, UT). The administration of the Concept Retrieval Technique included an instruction, requiring students to write down all concepts or ideas about a topic using only keywords or bullet points, with the directive to avoid writing in full sentences. The instructions were: “*Please write down all the concepts or ideas you have about the topic: Acids*” followed by the statement: “*Please do not write in full sentences, only keywords or bullet points will do.*” Students were given five minutes to complete the test and were required to fill in ten concepts that they could freely recall concerning the given topic.

After the administration of the Concept Retrieval Technique, two raters independently marked the responses with the help of the target word list. One mark was awarded for each correctly identified concept and a final score for each student was thereby generated. The final scores generated by each rater were compared for consistency and used to determine the overall reliability of scoring processes utilised by the Concept Retrieval Technique. See Figure 3.2 for an example of the scoring procedure.

Class	Index	Gender	Age	CRT4Qi1	CRT4Qi2	CRT4Qi3	CRT4Qi4	CRT4Qi5	CRT4Qi6	CRT4Qi7	CRT4Qi8	CRT4Qi9	CRT4Qi10
2RSP	27	Male	14	Acids contain hydrogen	Acid can break apart in water	More hydrogen ions found in 1000dg/m3 of water = Stronger Acid	Less hydrogen ions found in 1000dg/m3 of water = Weaker Acids	All acids are aqueous	Acids are corrosive and can "eat" up substances	Extremely low pH level	The lower the pH level, the higher acidity	The higher the pH level, the lower the acidity	Concentration of hydrogen ions determine the acidity
2RSP	13	Female	14	Some are corrosive	Toxic	Some are explosive	Concentration of an acid does not depend of number of Hydrogen atoms	Acids can be alkali	Acids can be acidic	Most acids are toxic	Acids should be used with the right equipment	Acids can cause death	Acids should not be handled with bare hands
2RSP	17	Female	15	hydrogen	amino acid	pH value	hydrochloric acid	diluted acid	concentrated acid	sulfuric acid	stomach	saliva	corrosive
2RSP	30	Male	13	Acids must contain H <sup>+</sup> ions <b>1 mark</b>	Difference in corrosiveness <b>1 mark</b>	Different pH levels	Dissolves in water <b>1 mark</b>	Aqueous	Acidic, Neutral, Alkaline	Concentrated <b>1 mark</b>	Strength of acid depends on how much it dissolves in water	Produces H <sup>+</sup> ions when dissolved in water	Hydrochloric Acid
2RSP	40	Male	14	not all acids are corrosive	produce hydrogen, h <sup>+</sup>	dissociate hydrogen partially from compounds is less acidic	measured in ph	some are harmful to health	some are helpful to life needs	can be found in carbonated drinks	dissociates hydrogen from compound completely is highly acidic	appearance does not affect the acidic level	can easily remove stains like rust
2CMT	21	1.00	14	Weak acids can be concentrated	Acidity	Strong acids can be less concentrated	Concentration of acids are determined by hydrogen molecules	Acids can be both harmful and helpful	Sulfuric acid is the strongest acid	Acids can be dangerous	Acid in lemon juice can produce electricity	Some acids are toxic	Hydrochloric acid is in our stomach

Figure 3.2. An example of the scoring procedures as part of Study 1.

## Analysis

Inter-rater agreement was established by computing Cohen's kappa ( $\kappa$ ) using the Statistical Package for the Social Sciences (SPSS), version 21. The value of the kappa can range from -1.0 to 1.0. According to Landis and Koch (1977), the value of  $\kappa$  can be categorized into six different interpretations as shown in Table 3.2. Therefore, a cut-off value of  $>.80$  is considered indicative of almost perfect agreement (Landis & Koch, 1977).

Table 3.2

*Interpretations of the values of Cohen's kappa,  $\kappa$*

Cohen's kappa, $\kappa$	Interpretation
$< 0$	Poor agreement
.01 to .20	Slight agreement
.21 to .40	Fair agreement
.41 to .60	Moderate agreement
.61 to .80	Substantial agreement
.81 to 1.00	Almost perfect agreement

## Results and Discussion

The results of the reliability analysis concerning what students could recall about the topic of “acids”, revealed high inter-rater agreement between the two independent raters. Kappa was equal to  $\kappa = .85$  demonstrating almost perfect agreement. In fact, only nine disagreements (out of a total of 73 pairs of independently rated scores) that arose as a result of the first (of two) raters miscounting the number of correct concepts in the target word list. This preliminary outcome suggests that having only to look out for concept words (listed on the target word list) in the responses and assigning one mark is a relatively straightforward procedure, with little error.

### 3.3 Study 2. Stability of Inter-rater agreement over subjects, age groups and raters

In study 2, we examined how stable kappa is when the Concept Retrieval Technique was administered for different school subjects and age groups and by different raters (i.e., teachers). If sufficiently high, this would be a strong indicator of how generalizable our procedure is for measuring conceptual knowledge. To this end, the Concept Retrieval Technique was administered in three different settings including a primary school science class (topic: properties of light), a secondary school science class (topic: diffusion) and a secondary geography class (topic: tectonic plates). We then generated the agreement score and examined whether the kappa values differed between the three administrations of the test. It was determined that if the kappa were of comparable value across the three samples, this would be interpreted as evidence for the generalizability of this test to measuring (conceptual) knowledge.

## Method

### Participants

In this study, 33 primary school science students from governmental schools in Singapore participated (46% female, age = 9 years,  $SD = .52$ ), 62 secondary school science students (45% female, age = 13 years,  $SD = .28$ ), and 116 secondary school geography students (53% female, age = 14 years,  $SD = .65$ ). They did not receive any compensation for their contribution.

### Materials

Concept Retrieval Techniques were developed, similar to that in Study 1. In addition, classroom teachers collaborated to develop target word lists for: (a) primary school science--topic: properties of light, (b) secondary school science--topic: diffusion, for the target word list, and (c) secondary geography--topic: tectonic plates, benefits of living near volcanoes. The target word lists used in all three topics are shown in Tables 3.3, 3.4 and 3.5 respectively.

Table 3.3

*The target word list of concepts used for the topic properties of light*

Admissible concepts	
particles	pass through materials
light sources	shadow
reflects off surfaces	

Table 3.4

*The target word list of concepts to be used for the topic diffusion*

Admissible concepts	
Net movement	Equilibrium
Constant, random motion	Passive process
Fast movement	Concentration gradient
Concentration	No energy
High to low concentration	

Table 3.5

*The target word list of concepts used for the topic benefits of living near volcanoes*

Admissible concepts	
fertile volcanic soil	geothermal energy
minerals	turbines
sulphur	electricity
precious stones	tourism
diamonds	Pompeii
building materials	

## Procedure

Similar to Study 1, each group of students, prior to the administration of the Concept Retrieval Technique, participated in an instructional lesson that addressed the concepts identified from the target word list. The duration of each lesson differed depending on the topic and age group. For the topic of diffusion, the duration was 30 minutes whilst for properties of light and tectonic plates it was 60 minutes. Each lesson consisted of a short lecture from the teacher followed by a period of self-study using written materials distributed by the teacher. After the lesson, the Concept Retrieval Technique was administered using the online survey platform Qualtrics®. The administration of the Concept Retrieval Technique involved an instruction, requiring students to write down all concepts or ideas about a topic in keywords or bullet points, avoiding writing full sentences. Students were given five minutes to complete each test and were required to fill in ten concepts for the primary and secondary school science classes whilst the secondary school geography class required 20 concepts, as the topic studied was more complex. After the administration of the Concept Retrieval Technique, two raters independently marked the responses with the help of the target word list. A total score was calculated for each student.

## Analysis

Inter-rater agreement was established by calculating Cohen's kappa.

## Results and Discussion

The kappa values for the three subjects, with different raters (i.e., teachers), yielded similarly high values as in Study 1. See Table 3.6 for the results. Inspecting the kappa values for the three samples, involving different age groups, subject domains and raters, one can see that the inter-rater agreement was high across all three data sets. All kappa's were  $\kappa > .85$  (mean kappa  $\kappa = .92$ ) and for primary school science there was even 100% agreement between raters. Furthermore, the three  $p$ -values were also found to be significant suggesting that the agreements in scores provided by the independent raters did not occur by chance. Overall, this outcome suggests that identifying and scoring concepts in the response texts is a straightforward and reliable affair.



Table 3.6

*The values of Cohen's  $\kappa$  generated for the three different samples*

Sample	<i>n</i>	$\kappa$	<i>p</i>
Primary school Science	33	1.00	<.001
Secondary school Science	62	.90	< .001
Secondary school Geography	116	.85	< .001

*Note:* The total number of students and kappa values for the reliability studies conducted across three different subjects and age groups.

### 3.4 Study 3. Scoring students' responses in full sentences

When the Concept Retrieval Technique was first administered to measure students' conceptual knowledge on the topic *Diffusion*, it was observed that although the students were given instructions to write down only keywords or bullet points, some students still wrote in full sentences. This may be because the Concept Retrieval Technique resembles the typical open-ended questions that students were accustomed to and therefore, found it unfamiliar to have to write only keywords or bullet points. Figure 3.3 below demonstrates an example of a student writing in full sentences, despite a reminder not to, to describe the relevant concepts.

Please write down **ALL** the concepts or ideas you have about the following topic:

**DIFFUSION**

**Important:** Try not to write full sentences, just keywords or bullet points will do.

- the cells that surround the stoma controls the size of the stoma
- the higher concentration of something in a place will go to a place where there is lower concentration
- diffusion occurs until equilibrium
- the cells are able to control the amount of carbon dioxide and oxygen in a leaf
- two places will eventually have the same amount of concentration of something
- nil
- nil
- nil
- nil
- nil

>>

*Figure 3.3.* An example of a student writing in full sentences for the Concept Retrieval Technique.

If raters of the Concept Retrieval Technique had to read through all of such responses before giving a score, the scoring procedure would be no different from that of open-ended questions, which have often been found to have low interrater agreement. Since the Concept Retrieval Technique was proposed to have an objective scoring procedure such that raters only look out for the correctly mentioned concepts based on a predetermined target word list of concepts, the purpose of this study was to compare whether there were any differences between (1) having to read and understand the full sentences that students have written and then provide a score, and (2) having to only look out for the correctly mentioned concepts that were representative of the underlying concepts leading to an entire semantic network of related concepts.

## Method

### Participants

A total of 62 students (28 females and 34 males) with a mean age of 12.95 years ( $SD = .28$ ) were involved in a session where the Concept Retrieval Technique was administered. As for the scoring of the responses, two independent raters were involved in the scoring of the Concept Retrieval Technique.

### Materials

The Concept Retrieval Technique was administered via an online survey software, Qualtrics®. The instructions involved in the Concept Retrieval Technique were “*Please write down all the concepts or ideas you have about the following topic: (e.g., Diffusion)*” followed by a reminder “*Do not write in full sentences; only keywords or bullet points will do.*” The Concept Retrieval Technique was administered at three different time intervals during a session to measure students’ conceptual knowledge regarding a secondary school science topic *Diffusion*. In each administration of the Concept Retrieval Technique, the students were asked to fill in 10 concepts for the topic in question.

### Procedure

To determine whether there were any differences between having to read and interpret the full sentences or identifying only the correct concepts in the responses, two raters independently scored the responses of the Concept Retrieval Technique twice. In the first round of scoring, the raters were asked to read through all responses, interpret them and then based on their judgment, assign one mark to the correct responses. In the second round of scoring, the raters were asked to

only look out for the correct concepts based on the predetermined target word list (see Table 3.7). This scoring method has to be done without having the raters to read or interpret any responses. As soon as the raters identify a correct concept, they can directly assign one mark for that concept.

### Analysis

Three independent *t*-tests were carried out to determine whether there were any significant differences in mean scores between the two scoring methods. All analyses were two-tailed and a *p*-value below .05 was considered statistically significant.

### Results and Discussion

A pair of independent raters scored the three sets of the Concept Retrieval Technique twice, the first time having to read and interpret the full responses given and the second time to look out for only the correct concepts based on the predetermined target word list. In total, the same pair of raters had to score the three sets of Concept Retrieval Technique twice with both scoring methods. The three sets of Concept Retrieval Technique scores were analysed with independent *t*-test comparisons and the results showed that there were no significant differences between the two scoring methods. The results to the three sets of scores can be found in Table 3.7.

Table 3.7

*The t-test results comparing the mean scores between the two scoring methods*

CRT	<i>t</i>	<i>df</i>	<i>p</i>
CRT 1	-1.81	98.07	.074
CRT 2	-1.63	105.39	.107
CRT 3	.01	107.76	.994

\**p* < .01

So, it can be concluded that the second scoring method of identifying correct concepts was as effective as having to read through and interpret the full responses from students. Although students were asked to respond to the Concept Retrieval Technique in only keywords or bullet points, it might also be interesting to consider the implications of students responding in full sentences like in open-ended questions. Because open-ended questions are commonly known to assess conceptual understanding (Schuwirth & Van der Vleuten, 2011), when students responded

in full sentences, it demonstrates that they have developed an understanding about the concept relevant to the topic in question. Take for example: a full sentence response from one of the students is “difference in concentration between two regions is concentration gradient” and the correct concept as determined by the target word list is *concentration gradient*. According to the learning materials, since concentration gradient is indeed defined as the difference in concentration between two regions, this full sentence response provides a clear indication that the student has understood the concept and is able to describe the meaning behind the concept *concentration gradient*. While the objective of the Concept Retrieval Technique is not to assess conceptual understanding, the result of this study suggests that there is more to the Concept Retrieval Technique than just a simple measure of what examining how many correct concepts students can recall.

### 3.5 Summary of findings

The objective of studies 1, 2 and 3 was to establish reliability evidence for a new measure of conceptual knowledge that requires the test-takers to freely recall relevant concepts about a specific topic; hence the name Concept Retrieval Technique. In Study 1, the reliability of the Concept Retrieval Technique was examined by comparing the degree of agreement between two independent raters using a predetermined target word list of admissible concepts to score the students’ responses. The results revealed that there was high agreement between both raters suggesting that they were able to consistently assign a correct mark for each recalled concept corresponding with the target word list. In Study 2, the reliability analysis was extended by examining how consistent, or generalizable, the inter-rater agreement was with different raters and across different subject domains (ranging from science to geography) and age groups (primary school and secondary school in Singapore and Australia). In Study 3, two scoring approaches were compared. The first scoring method involved getting raters to read and interpret the full responses by students. The second method was a simplified approach that entailed spotting the correct concepts without trying to make sense of the entire response. The results showed that there were no significant differences between the scores of the detailed and simplified scoring method. However, the main difference between both methods was that the simplified method, requiring spotting of target concept words, resulted in a substantially higher interrater agreement. The results revealed that across all conditions, high interrater agreement could be achieved, which attests to the overall consistency in generating the Concept Retrieval Technique scores.





# 4

Chapter

## The Validity of the Concept Retrieval Technique

### Abstract

After having established that the Concept Retrieval Technique can be reliably scored across different subject domains, age groups, and with different raters, the aim of the studies presented in this chapter were to establish the validity evidence for the Concept Retrieval Technique as a precise measure of conceptual knowledge in an educational environment. Study 4 was concerned with establishing the convergent validity of the CRT ( $N = 55$ ). The correlation between the Concept Retrieval Technique scores and scores on essay-type items was  $r = .69$ . In Study 5, an experiment was conducted to determine the construct validity of the Concept Retrieval Technique ( $N = 45$ ). Participants either acquired, or did not acquire, new knowledge of a particular topic and this manipulation was reflected in their Concept Retrieval Technique scores. The results of the studies suggest that the Concept Retrieval Technique is a valid instrument to measure conceptual knowledge.

#### 4.1 Validity evidence for the Concept Retrieval Technique

Typically, when testing a new instrument, the first piece of validity evidence results from an examination of how well the scores obtained are associated with the scores from an existing measure. This is referred to as “convergent validity” (Campbell & Fiske, 1959). Convergent validity refers to the extent to which a measure correlates with other measures from other tests that are theoretically predicted to correlate with. In line with this approach, for Study 4 we correlated the scores obtained from the Concept Retrieval Technique with scores obtained from a test utilising essay-type items about the same topic. Both instruments are aimed at measuring students’ conceptual knowledge of the periodic table. If the correlation would be positive, sizable and significant, we considered it as evidence, that the Concept Retrieval Technique is capable of capturing the conceptual knowledge of a test-taker at a similar level as an essay-type test with lengthier written responses and explanations.

However, it should be noted that this form of validity evidence has recently received some criticism because all measurements that are intended to measure similar things are, to some degree, associated with each other (even if they are taken at different points in time as it was the case in our study). According to Borsboom et al. (2004), if one is concerned with establishing the validity of a new measure, what needs to be demonstrated is that manipulation of the (non-observable) attribute being measured results in changes in the scores of the measure. This approach is, however, impractical when measures of personality traits are involved; it is difficult to manipulate traits in people. For instance, the construct validity of an intelligence test would be difficult to

ascertain following Borsboom's suggestion because intelligence is supposed to be highly insensitive for attempts at change. However, when it comes to the Concept Retrieval Technique, this approach seems to be a robust means to establish its construct validity because someone's level of knowledge can be increased by means of instruction. Therefore, an experiment was conducted in which we manipulated the amount of knowledge students were able to acquire and examined whether this manipulation could be measured with the Concept Retrieval Technique.

In other words, a more convincing form of validity arises if deliberate manipulations of the construct the instrument is intended to measure, result in *changes* in the scores measured by the instrument. Therefore, in Study 5, we manipulated the amount of knowledge students were able to acquire. A treatment group received relevant information, whereas irrelevant information was provided for a control group. We hypothesized that if this manipulation was reflected in the Concept Retrieval Technique scores, it can be considered evidence for both its ability to discriminate in measuring conceptual knowledge and, concomitantly, of its construct validity.

When questioning the validity of an assessment, it is equivalent to asking whether relevant information can be gathered from the test to make inferences about students' performance. While validity and reliability are two different concepts used to assess the quality of an assessment, these two concepts are closely associated with each other. For instance, a test cannot be valid unless it is reliable but at the same time, the reliability of a test does not necessarily depend on its validity.

## 4.2 Study 4. Convergent validity of the Concept Retrieval Technique

Study 4 was concerned with establishing the convergent validity of the Concept Retrieval Technique. This was done by correlating the Concept Retrieval Technique scores with scores obtained from essay-type questions. If the Concept Retrieval Technique is a valid instrument that is capable of measuring conceptual knowledge, the scores should show a significant and positive correlation with the scores obtained from essay-type items that measure the same topic (i.e., the arrangement of elements in the periodic table).

### Method

#### Participants

Fifty-five secondary school science students participated from an all-male secondary school in Australia. Their average age was 15 years ( $SD = .52$ ). Participation was voluntary and students did not receive any compensation for their contribution.



## Materials

**Exam Scores.** Participants' exam scores were obtained about the topic periodic table. There were two open-ended items of which the scores were aggregated. See Figure 4.1 for an overview of the items and marking scheme. The inter-rater agreement, expressed by the kappa, was  $\kappa = .72$ .

**Concept Retrieval Technique.** A Concept Retrieval Technique on the periodic table was administered one week before the science exam during which the open-ended items were administered and followed a similar approach to that in both previous studies. Teachers collaborated with researchers to generate the target word list containing all admissible concepts for this topic. See Table , for an overview of the target word list for this topic. The inter-rater agreement for this CRT administration was  $\kappa = .85$ .

### Question 26 (4 Marks)

For the given table showing atomic numbers and mass numbers of elements.

- What is the largest number of electrons that can fit into the first shell of each of the atoms in the table?
- Which element in the table has 8 protons in the nucleus of its atoms?
- What is the electron arrangement of a sodium atom?
- Explain why the sodium atom has no electrical charge.

### Marking Scheme

	Criteria	Mark
26(a)	• Identifies the correct answer	1
26(b)	• Identifies the correct element	1
26(c)	• Provides the correct electron arrangement	1
26(d)	• Provides the correct explanation	1

### Question 28 (5 Marks)

A student investigated heating metal carbonates. The student set up an experiment and tabulated the data.

- Use the correct answer from the box to complete the sentence.

black	green	white
-------	-------	-------

The color of copper oxide is .....

- Solution A is used to test for carbon dioxide. Carbon dioxide turns Solution A cloudy. What is the name of solution A?
- Most metal carbonates produce the metal oxide and carbon dioxide when heated. What is this reaction known as?
- Use the information from the table given and state TWO reasons why potassium carbonate did not react.

### Marking Scheme

	Criteria	Mark
28(a)	• Correctly identifies the colour	1
28(b)	• Provides correct answer	1
28(c)	• Identifies the correct chemical reaction	1
28(d)	• Correctly states TWO reasons • Correctly states ONE reason	1-2

Figure 4.1. Open-ended questions and marking schemes used within the assessment of the general science examination.

Table 4.1

*The target word list used for scoring the topic The Periodic Table*

<u>Admissible concepts</u>	
Atomic number	Atomic mass
Electrons	Reaction
Formula	Element
Protons	Atom
Element example	Shells

### Procedure

Prior to the administration of the Concept Retrieval Technique, students had studied the periodic table as part of their science project titled “The Mighty Atom”. The project was five weeks in duration and was completed prior to their yearly general science examination. In this project, students worked in groups to classify newly discovered elements and justify their possible position in the periodic table. Each group was given a (hypothetical) newly discovered element and they were required to create a presentation that distinguished their element based on their understanding of neutrons, protons, electrons and the periodic table. Students were provided with a range of learning materials (including tutorials and lectures) to assist them in acquiring conceptual knowledge on the topic-at-hand. Also, they were given self-study time in their groups to research and in the preparation for their presentation.

The Concept Retrieval Technique was administered one week prior to the yearly general science examination and, unlike the previous studies, there was no specific instructional lesson prior to the administration of the Concept Retrieval Technique. Subsequently, the Concept Retrieval Technique for “the periodic table” was aimed at the activation and elaboration of prior knowledge obtained throughout their participation in the project. After administration of the Concept Retrieval Technique, two raters independently marked the responses with help of the target word list. The scores of the two questions relevant to the topic-at-hand were retrieved from this general science examination and their mean score was correlated with the Concept Retrieval Technique scores.

### Analysis

Inter-rater agreement was established by computing Cohen’s kappa using the Statistical Package for the Social Sciences (SPSS), version 23. A Pearson’s product-moment correlation

coefficient ( $r$ ) was calculated to examine the strength of association between the exam scores and the Concept Retrieval Technique Scores. According to Cohen (1988), the strength of the correlation can be interpreted as shown in Table 4.2.

Table 4.2

*Interpretations of the Pearson Product-Moment Correlation Coefficient,  $r$*

Pearson product-moment correlation coefficient, $r$	Interpretation
$\pm .10$ to $\pm .29$	Small
$\pm .30$ to $\pm .49$	Medium
$\pm .50$ to $\pm 1.0$	Large

## Results and Discussion

The results of the reliability analysis concerning both assessment measures, revealed interesting inter-rater agreements between the two independent raters. Firstly, on the open-ended questions within the general science examination the kappa was equal to  $\kappa = .72$ , demonstrating substantial agreement. Secondly, on the Concept Retrieval Technique the kappa was equal to  $\kappa = .85$ , demonstrating almost perfect agreement. The analysis of the kappa values provides strong evidence of the issues with reliability that occur in assessments that use open-ended questions. Moreover, the results of the analysis revealed a large positive correlation between the exam scores and the Concept Retrieval Technique scores:  $r = .69$ ,  $p = .001$ . This outcome suggests that both measures are relatively highly associated; students who scored relatively high on the Concept Retrieval Technique also scored relatively high on the exam (and vice versa). This finding suggests that the Concept Retrieval Technique has considerable convergent validity. However, as stated before, a correlation between two test scores cannot provide conclusive evidence regarding the validity of a measure as it merely suggests that both are related. To address this potential limitation Study 5 was conducted.

### 4.4 Study 5. Construct validity of the Concept Retrieval Technique

To explore the construct validity of the Concept Retrieval Technique, a randomised controlled experiment was conducted. The experiment was set up as a problem-based learning exercise, during which participants received a problem about infectious diseases, discussed it in pairs and engaged in independent self-study (Schmidt, Van Der Molen, Te Winkel, & Wijnen, 2009). Participants were randomly assigned to two conditions, a control group and a treatment

group, which enabled us to manipulate the participants' opportunity to acquire knowledge regarding the topic at hand (i.e., infectious diseases). Only the treatment group received relevant information, whereas the control group did not. The Concept Retrieval Technique was administered at three points in time: at the start of the session, after presentation of the problem and discussion, and at the end after studying the respective texts.

We hypothesised that students would not gain a significant amount of knowledge during the first phase when the problem was presented and discussed. (Although some minor variation in students' prior knowledge regarding the topic was expected.) However, during the second phases we hypothesised to observe a significant increase in the Concept Retrieval Technique score for the treatment group. This was to be expected because only the treatment group received the text about infectious diseases, whereas the control group studied an irrelevant text about evolution, thus they were not in the position to acquire concepts with regards to infectious diseases. If the Concept Retrieval Technique is a valid measure of a person's semantic knowledge, this must be reflected in a significant higher Concept Retrieval Technique score for the treatment group.

## Method

### Participants

Forty-five secondary school science students participated from the same school as in Study 4 (100% male). The students were randomly selected and assigned to either a control or treatment group. Their average age was 13 years ( $SD = .32$ ). Participation was voluntary and students did not receive any compensation for their contribution.

### Materials

Concept Retrieval Techniques were developed, similar to that in three previous studies. Teachers in collaboration with researchers generated a target word list for the topic "infectious diseases". Subsequently, three Concept Retrieval Techniques were administered about the topic-at-hand. Inter-rater agreements for these tests were computed to show  $\kappa = .91$ ,  $\kappa = .94$  and  $\kappa = .90$  for each test respectively. See Table 4.3 for the target word list.

Table 4.3

*The target word list used for scoring the topic infectious diseases*

<u>Admissible concepts</u>	
Influenza	Contagious
Bacteria/Bacterial	Aids/Ebola
Protozoan/Sporozoan	Infection
Flagellates	Medicine
Blood	Immune
Vaccine/Vaccines	

### Procedure

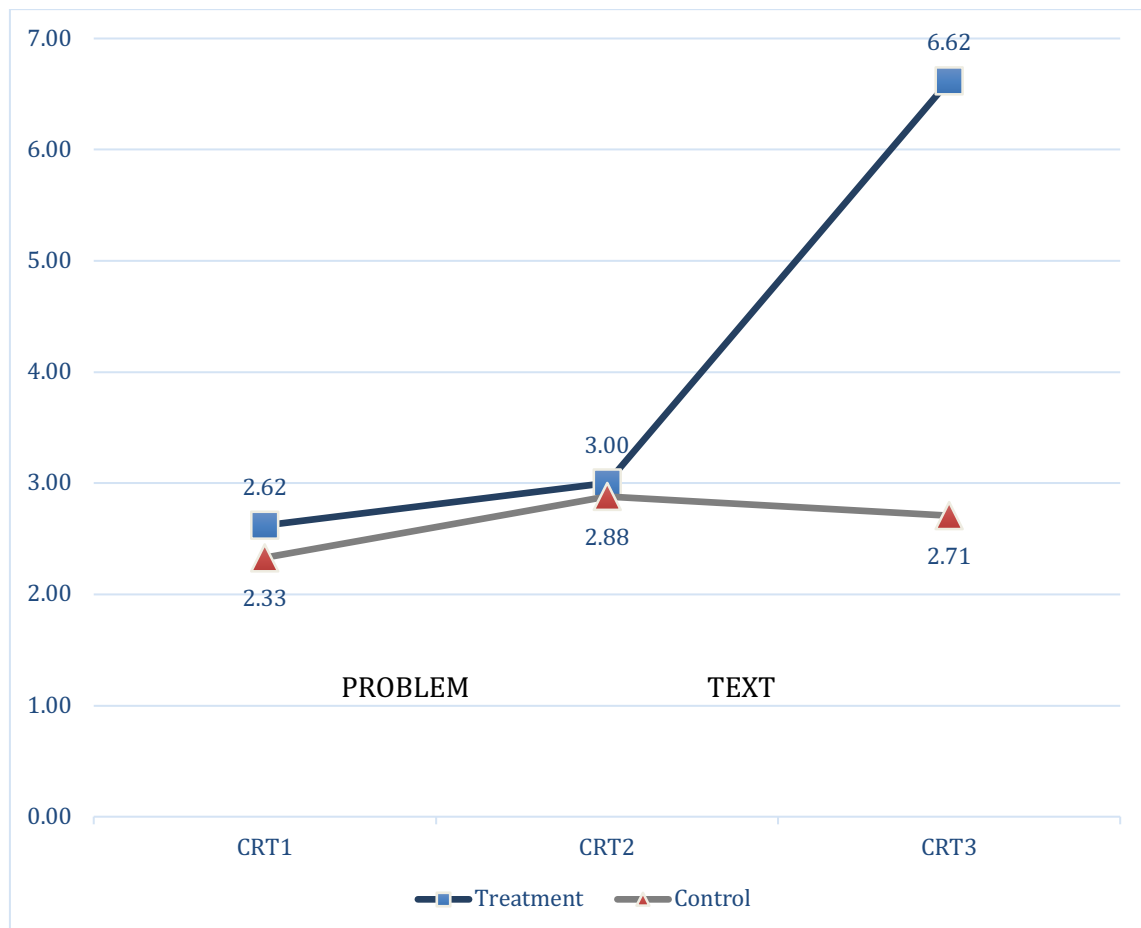
The duration of the experiment was 30 minutes. All participants were required to respond to the Concept Retrieval Technique at three points in time. The first administration was before the session commenced. Students were given their first Concept Retrieval Technique to measure the baseline knowledge of infectious disease of both groups. Participants then were presented with a problem about infectious diseases and discussed the problem in pairs. This took 10 minutes. The problem reads as follows: *“Influenza (Common Flu) was introduced into Australia by the early settlers. Why would the flu be deadlier to Aboriginal people than European Settlers?”* The purpose of this exercise was to activate and share prior knowledge, which would facilitate the processing of new information (Schmidt, Rotgans, & Yew, 2011). A second Concept Retrieval Technique was administered to measure the changes in conceptual knowledge as a result of sharing their prior knowledge in pairs. Subsequently, students engaged in individual self-study of their assigned textbook chapter from the *“CPO Focus on Life Science”* (Eddleman, 2007). The treatment group received a text that contained information relevant to the problem (topic: the microscopic world), whereas the control group received a nonrelated text (topic: evolution). Studying the text lasted for about 20 minutes. Finally, a third Concept Retrieval Technique was administered to show the learning gain of students from their engagement with their assigned text during the self-study period. After the administration of this experiment, two raters independently marked the responses of all three Concept Retrieval Techniques using the assigned target word list.

## Analysis

A 2\*3 repeated-measures ANOVA with a between-group factor (treatment vs. control) and Concept Retrieval Technique measurements (measurement 1, measurement 2 and measurement 3) as the within-group factor was conducted. With this analysis, we examined if there were significant differences between both groups in terms of their knowledge trajectories measured with Concept Retrieval Technique over the three measurement occasions.

## Results and Discussion

In this study, we manipulated the participants' opportunity to acquire knowledge regarding the topic of infectious diseases. Only the treatment group received relevant information regarding the topic. A visual overview of the results is depicted in Figure 4.2.



*Figure 4.2.* Mean Scores of the three Concept Retrieval Techniques. The control group received irrelevant information and the treatment received relevant information on the topic assessed by the Concept Retrieval Technique.

The results of the 2\*3 repeated-measures ANOVA suggest that there was a significant main effect: Wilk's  $\Lambda = .18$ ,  $F(2, 86) = 96.77$ ,  $p < .001$ ,  $\eta^2 = .82$  and more importantly, also a significant interaction effect: Wilk's  $\Lambda = .17$ ,  $F(2, 86) = 105.86$ ,  $p < .001$ ,  $\eta^2 = .83$ . This outcome suggests that there was a significant difference between the treatment and control group in terms of the Concept Retrieval Technique scores. See Table 4.4 for details.

Table 4.4

*Descriptive Statistics for the Experimental and Control Groups for the Three Concept Retrieval Techniques*

Tests	Control			Treatment		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
CRT1	24	2.33	1.40	21	2.62	0.86
CRT2	24	2.88	1.30	21	3.00	1.00
CRT3	24	2.71	1.16	21	6.62	1.32

*Note:* Mean values and standard deviations for the three concept Retrieval Techniques administered under experimental conditions (control: irrelevant text and treatment: relevant text regarding the assessment).

Planned post-hoc pairwise comparisons further suggest that there were no significant differences between the first two measurements of the Concept Retrieval Technique; both groups possessed the same amount of knowledge, prior and post the problem discussion. However, this was different after both groups had an opportunity to study their respective texts. In-line with our manipulation, only the treatment group (who studied the text about infectious diseases) resulted in significantly higher Concept Retrieval Technique scores as compared to the control group (who studied an irrelevant text) ( $F(1, 44) = 111.75$ ,  $p < .001$ ,  $\eta^2 = .72$ ). In short, manipulation of the target attribute (knowledge regarding infectious diseases) resulted in a corresponding change in Concept Retrieval Technique scores, which suggest that this instrument is capable of measuring the conceptual knowledge of the participants for the topic in question.

#### 4.5 Summary of findings

The objective of the studies 4 and 5 was established through an examination of its convergent and construct validity. In study 4, this was done by first correlating the scores of the Concept Retrieval Technique with the scores of open-ended items on a conventional exam about the same

topic (i.e., the periodic table). The results revealed a relatively large positive correlation between the scores of both measurements, explaining almost 50% of the variance. In Study 5, we manipulated the amount of knowledge students could acquire about a particular topic, namely infectious disease. A treatment and a control group received an unequal amount of knowledge about this topic. To that end, first, baseline knowledge of students was assessed using a Concept Retrieval Technique at the start of the experiment. Second, concept retrieval was measured after students were given the opportunity to participate in a problem discussion on the topic. Third, after the treatment group received a text about the microscopic world whereas the control group read a text about evolution, a third Concept Retrieval Technique was taken.

The outcome of the experiment revealed that the Concept Retrieval Technique was capable of picking up changes in participants' conceptual knowledge regarding the topic in question; in short, the manipulation of the criterion was reflected in the Concept Retrieval Technique scores. Taken together, these findings suggest that the Concept Retrieval Technique scores closely resemble the conceptual knowledge participants had regarding the topic in question.

#### **4.6 Discussion of key findings**

What are the discriminating features of the Concept Retrieval Technique, that sets it apart from other measures of student learning? The first is that we have shown it to be a highly reliable measure of knowledge acquired by students, more reliable than most of its competitors. Concept mapping and open-ended questions, as used in the day-to-day practice of education, tend to be reliable only to a limited extent, making decisions about student performance sometimes a murky process. Although it is possible to construct MCQ examinations that are highly reliable, its implementation often requires time-consuming production of large numbers of items and subsequent extensive item analysis. Second, construction and administration of the Concept Retrieval Technique is simple and does not require much teacher time, with the exception of the construction of a target word list against which the responses of the students need to be scored. Administration of a Concept Retrieval Technique usually takes less than five minutes, enabling frequent testing without much disruption to the ongoing learning.

We foresee that the construction of a target word list, based on the subject matter that students are supposed to study, and the scoring of their responses can be easily automatized. An implication is that information about students' performance will become available almost in real time, as the learning proceeds. Yew, Chng and Schmidt (2011) have used this feature of the Concept Retrieval Technique while studying knowledge acquisition in a problem-based



curriculum. They measured knowledge at four points in time; before a problem was presented to students, after the students had discussed the problem, after their individual study and after a second round of discussion of the problem. They were able to demonstrate that knowledge acquisition in problem-based learning is cumulative, and that each activity both helps consolidating previously acquired knowledge and adds to the knowledge base of students. In addition, learning as measured by the Concept Retrieval Technique predicted final achievement. Rotgans and Schmidt (2014) were able to show that during an active learning sequence knowledge almost linearly increased as a function of time.

The idea of a Concept Retrieval Technique is based on an explicit theory of how declarative knowledge is represented in long-term memory: concepts are nodes in a semantic network kept together by a more or less extensive assemblage of links. The more concepts, and the more links, the more coherent the network. Activating one of those concepts through the instruction in a Concept Retrieval Technique, will launch a process of spreading activation through that network thereby bringing concepts into consciousness as a function of network coherence.

Finally, the construction of an online automated marking process is also essential. The strengths of the Concept Retrieval Technique lie in its reliability and speed of application. Hence, an automated marking process will provide teachers with a unique opportunity to reliably assess students by accessing real-time data within the lesson and thereby maximising the learning by making appropriate and more immediate adjustments to their teaching. The development of this software would dramatically increase the rate of the marking and immediacy of feedback when compared to doing so manually. Unlike human raters, a computerised system would be very efficient at identifying concepts quickly, however, the ability to make decisions concerning spelling mistakes and the use of alternative spelling and words may be problematic for an automated program. Therefore, an automated Concept Retrieval Technique must be tested in variety of conditions, including within educational settings that utilise active-learning pedagogies, and must be able to discern mistakes and alternate spelling and words contained in student responses.



# 5

Chapter

## **Designing an Automated Concept Retrieval Technique**

## 5.1 Introduction

The aim of this chapter is to explore the identified limitations with automated assessment and discuss the reasons why the automated Concept Retrieval Technique is a reliable and valid assessment measure in light of these limitations. The strengths of the Concept Retrieval Technique lie in its speed of application and high levels of human-machine agreement and test-retest reliability. Discussion within this chapter will focus on the nature and feasibility of automated assessment, addressing the current issues with automated essay scoring systems. Also, the purpose for creating an automated Concept Retrieval Technique will be discussed and the possible benefits explored. Such as providing teachers with the ability to obtain valid measures of student conceptual knowledge and understanding, acquired in real-time from learning episodes. As a result, teachers can make immediate adjustments to teaching, targeted to specific student levels of cognition and understanding. This chapter will include a review of the literature outlining the nature and feasibility of automated assessment, a detailed background of the design process for construction of the automated Concept Retrieval Technique (Version 1) and findings from its initial trial will be presented. Limitations will be discussed and will inform the development of future versions (i.e., Version 2 and 3 that are discussed in Chapter 7 and 8).

## 5.2 The nature of automated assessment in an educational context

Over the past decade, there have been considerable technological advances that have motivated a vast increase in the use of automated assessment in education. These technological advancements have enabled a more flexible, efficient and personalised assessment experience for all stakeholders (Bejar, Williamson, & Mislevy, 2006; Siddiqi, Harrison, & Siddiqi, 2010; Zhang, 2013). As a result, this has caused a conceptual shift in traditional assessment paradigms, moving away from exclusively using summative assessment and broadening the application of formative assessment as an appropriate measure of student knowledge. The high cost of human scoring and reliability issues, has contributed to an increased use of formative assessment. Consequently, enabling the efficient collection of assessment data within learning episodes to provide instant feedback, based on a learners' individual learning needs and results, informing future learning opportunities (Attali, Lewis, & Steier, 2013; Redecker & Johannessen, 2013). In addition, the need to provide a more personalised learning experience based on formative assessment data, has resulted in the development of more sophisticated automated assessment algorithms. These advancements have the potential to further revolutionise assessment paradigms, that provide more personalized and targeted learning experiences for all students. Specifically, during inquiry

instruction in active learning environments, using ongoing assessment feedback to provide adaptive guidance that is agile and responsive to student learning needs (Azevedo & Bernard, 1995; Krause, Stark, & Mandl, 2009).

The use of automated assessment has a number of key advantages, such as paperless test distribution and data collection, increased assessment efficiency, rapid feedback and the opportunity for machine-scoring. The major catalyst for the increased use of automated assessment in education has been machine-scoring, which has the potential to dramatically reduce the time and cost associated with written assessments and increase reliability (Dikli, 2006). Despite these benefits, there is still scepticism concerning the application of machine-scoring especially with high stakes assessments. The ultimate validity criterion for machine-scoring is in its ability to emulate human raters with similar means and standard deviations, mimicking the interpretive powers of human experts in consistently aligning scores to scoring rubrics (Attali, 2015; Bejar et al., 2006; Bridgeman, Trapani, & Attali, 2012; Cushing Weigle, 2010). The reliability, construct and predictive validity of machine-scoring have been thoroughly explored, these results suggest that automated assessment performs as well or better than human raters (Liu, Rios, Heilman, Gerard, & Linn, 2016; Shermis, Page, & Keith, 2002). However, criticism still exists with automated assessment and these perceived issues have limited the use of machine-scoring to low stakes testing situations, minimising validity issues by using multiple-choice questions (Kersting, Sherin, & Stigler, 2014; Shermis, 2014; Zhang, 2013).

By contrast, human scoring is not excluded from its own shortcomings, including inter-rater reliability issues, rater biases, rater drift and the cognitive limitations of raters (Bejar, 2011; Bennett & Bejar, 1998; Cushing Weigle, 2010; Zhang, 2013). Machine-scoring has the potential to overcome obvious issues and limitations with human scoring, such as greater consistency in scoring. Furthermore, consistent interpretation and execution of a scoring criteria are always identical and replicable, yielding perfect re-test reliability in machine-scored assessments (Bejar et al., 2006; Cushing Weigle, 2010).

### **5.3 The feasibility of machine-scoring in an educational context**

Regardless of whether an assessment is scored by human raters or by a machine-scoring mechanism, the goal of assessment is to ensure that the construct is appropriately represented in the final scores of the assessment process (Bejar et al., 2006). The feasibility of machine-scoring is dependent on factors such as the content domain, complexity of the task, type of assessment item (e.g., MCQs vs. essay type questions), variables within the assessment task design (e.g., Time,

complexity and learning environment) and the responses used to build the automated scoring model (Liu et al., 2016).

Historically, the type of assessment item has been a limiting factor in machine-scoring. Typically, conventional assessment formats have been administered (i.e., multiple-choice, true/false or open-ended questions), with MCQs being predominately used. MCQs are generally regarded as robust and reliable in machine-scoring (Birenbaum & Feldman, 1998; Schuwirth & Van der Vleuten, 2011). However, concerns are raised that MCQs only assess recognition of the correct answer often promoting guessing effects, rather than requiring the retrieval of knowledge (Bejar et al., 2006; Glass & Sinha, 2013; Nicol, 2007; Rodriguez, 2005). Given that MCQs are not suitable in assessing a deep understanding of the learning concepts, attention should be focused on the validity of automating the scoring of open-ended or free-text responses (Noorbehbahani & Kardan, 2011).

Scoring knowledge retrieval by free-text assessment items, such as short-answer or essay type assessment items is a challenging and complex process. Human raters are required to develop a common interpretation of the scoring criteria. This is achieved by pilot marking sessions that help raters maintain reliable judgements during the scoring process. Furthermore, essay samples are used as exemplars within the scoring benchmarks of the rubric, ensuring consistent agreement between raters is maintained and reducing the occurrence of rater drift (Attali & Powers, 2009). Despite these quality assurance procedures, the cost is considerable and issues with inter-rater and test-retest reliability still transpire. Progress has been made on improving machine-scoring techniques for free-text assessment items by the continually enhancements made to scoring algorithms. However, a common issue is the ability of the algorithm to understand and interpret free-text that do not conform to the programming criteria in essay responses. Furthermore, the semantic analysis of the text is often limited to short-answer responses, especially when compared with human raters (Attali, 2015). Specifically, for short-answer free-text responses of around one sentence, machine-scoring has also been shown to be at least as good as human markers (Attali et al., 2013; Attali & Powers, 2009; Bridgeman et al., 2012; Butcher & Jordan, 2010; Cushing Weigle, 2010; Kersting et al., 2014; Liu et al., 2016; Noorbehbahani & Kardan, 2011; Redecker & Johannessen, 2013; Shermis, 2014).

Aside from the identified issues with selection of assessment items in automated assessment, the effective design of the assessment environment is a key consideration critical in addressing the validity argument of machine-scoring. As with human scoring, the accuracy of automated scoring depends on several assessment design factors, including task clarity, quantity and complexity of data collected and well-designed assessment instructions (Bejar, 2011; Bejar et

al., 2006; Bennett & Bejar, 1998). An obvious difference between human and automated scoring is the explicitness, and therefore tractability, of the scoring (evidence identification) instructions. Instructions for human scoring, as part of large-scale assessments, consist of a set of guidelines called a “scoring rubric.” Despite training intended to minimize the individual variation in application of this rubric it is likely that each human grader implements the rubric in a slightly different way (Bejar et al., 2006).

#### **5.4 The Concept Retrieval Technique as an effective instrument for automated assessment**

Automated essay scoring systems, such as e-rater® (Attali & Burstein, 2006) have led the field in producing favourable results with machine-scoring. However, human raters have the ability to score writing according to quality, while automated essay scoring systems can only identify surface text quality, often limited to feedback on the mechanics of writing. The automated scoring system is used to determine elements such as average sentence length, variety of sentence type, the occurrence of topic-related vocabulary items, rather than measuring conceptual knowledge and understanding (Condon, 2013). As a result, automated assessment systems tend to omit features that cannot be easily computed, such as content, organization and development (Ben-Simon & Bennett, 2007). A study by Deane (2013), showed a high correlation between human and machine-scoring, revealing that the average quadratic-weighted kappa for automated essay scoring systems ranged from .60 to .84, in comparison to .61 to .85 for humans. Unfortunately, discrepancies predominately occur in assessing more abstract aspects of conceptual knowledge, especially using free-text responses. Often, in these situations human raters have the capacity to make judgments on the validity answers, whereas machine-scoring can only mark responses according to expert knowledge and programming rules applied to the program. Consequently, there are some remaining challenges that need to be addressed before it can be applied to the Concept Retrieval Technique.

The first challenge is concerned with the overuse of MCQs items in automated assessment. The issues with MCQs have been previously expressed, specifically the emphasis on answer recognition, rather than retrieval of knowledge and the construction of a response. This affirms the need for an automated Concept Retrieval Technique that has the potential to assess conceptual knowledge and understanding with the assistance of machine-scoring. Butcher and Jordan (2010) demonstrated that with short-answer free-text responses, machine-scoring produces consistent inter-rater reliability and strong validity with human raters. The Concept Retrieval Technique is

proposed as an effective instrument for automated assessment, as it addresses the retrieval deficiencies associated with MCQs and the validity issues identified with machine-scoring essay type questions. Given, the Concept Retrieval Technique only accepts short-answer free-text responses of around one sentence it is already an effective alternative to MCQs.

The second challenge, is the occurrence of students incorrectly spelling target word list concepts. Misspelled words are common in student writing and have been identified as one of the most frequent errors in student essays. Connors and Lunsford (1988) reported that approximately 25% of all errors found in a sample of 300 students' essays were misspelled words. Although humans can automatically identify the equivalence of a word, it is clear that spelling errors are a common feature of free-text responses and considerations need to be employed to address this issue. The opportunity for the automated Concept Retrieval Technique to access spell-checking submodules that can clean submitted assessment data is an option to address this issue. Flor and Fugati (2012) found that automatic detection and correction systems still have significant limitations especially in the use of domain-specific words. Their studies found that automated spell-checking programs did not correct more than 20% of the misspelled words in the text, and moreover that 20% of the suggestions by the correction program were wrong. This raised concerns regarding the inability of computers to infer the true meaning of misspelled words, the occurrence of spacing errors that could cause the algorithm to erroneously classify an incorrect new word and hinder the machine-scoring process (Flor & Fugati, 2012; Muhlenbach, Lallich, & Zighed, 2004).

The third challenge, which is universal to all assessments including human scoring, is the ability of the machine-scoring engine to effectively discern contrasting user responses to the responses listed in the target word list. The occurrence of diverse concepts that have a similar meaning presents a considerable challenge to the reliability in scoring with the automated Concept Retrieval Technique. Butcher and Jordan (2010) acknowledge that human raters have the potential to make interpretation decisions during assessment, which is not afforded to machine-scoring. Therefore, it is imperative that the automated Concept Retrieval Technique minimises the effect on the overall reliability due to the occurrence synonyms for target concepts.

Therefore, the following additions will be utilised to minimise the occurrence of the challenges stated above. Firstly, wildcard characters will be utilised to reduce the occurrence of issues with the use of plurals for target concepts. This will also address some issues with spelling that may arise. For instance, Figure 5.1 demonstrates how the use of wildcard characters preceding and proceeding the target word in the search engine will correctly score responses similar to a human rater. In the example, the target word is *cell* and the search engine will use wildcard

characters symbolised by the \* character. The wildcard means that any characters can be present preceding and proceeding the target word and a match will still be identified. In the example *blood cell* would be scored correctly due to the wildcard, however *ceel* would be scored as incorrect.

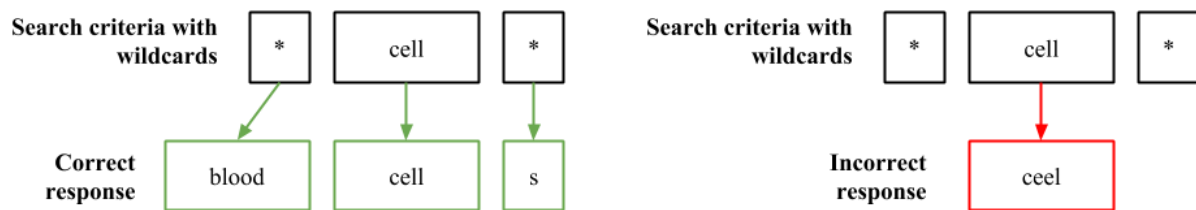


Figure 5.1. The use of wildcard characters in the scoring engine of the automated Concept Retrieval Technique to score responses.

Secondly, by providing the user with a visualisation detailing the frequency of target word list concepts in submitted responses. It is anticipated that common issues with synonyms or spelling will be diagnosed, affording users the opportunity to amend the target word list before scoring is undertaken. The visualisations will form a word cloud, which is a visual portrayal of words, with the more frequently occurring words appearing depicted in a larger font or different colours (Jayashankar & Sridaran, 2017; Ramsden & Bate, 2008). For instance, Figure 5.2 provides an example of a word cloud generated from student responses to the Concept Retrieval Technique for the topic *periodic table*. The data contained within this word cloud expose for the rater spelling issues with the concept *neutron* as it is misspelt consistently as *nueutrons*. It would be implied that these students would have the conceptual knowledge linking the concept *neutron* to the *periodic table* and therefore should receive a mark in the scoring process. Furthermore, plural issues with the concepts *reaction*, *neutron* and *atom* also exist and this provides the user with feedback concerning the target word list to be employed in the scoring process. They may enter *atom* in the target word list, knowing that concepts such as *atomic* and *atoms* will be scored correctly as a result of the wildcard feature. Finally, the word cloud also provides the user with feedback regarding the effectiveness of the learning episode. Consequently, if it is expected by the user that certain target concepts should be predominately represented in test-taker responses and they are absent. Then this provides the user the opportunity to address the confusion and assists with future planning of learning opportunities.



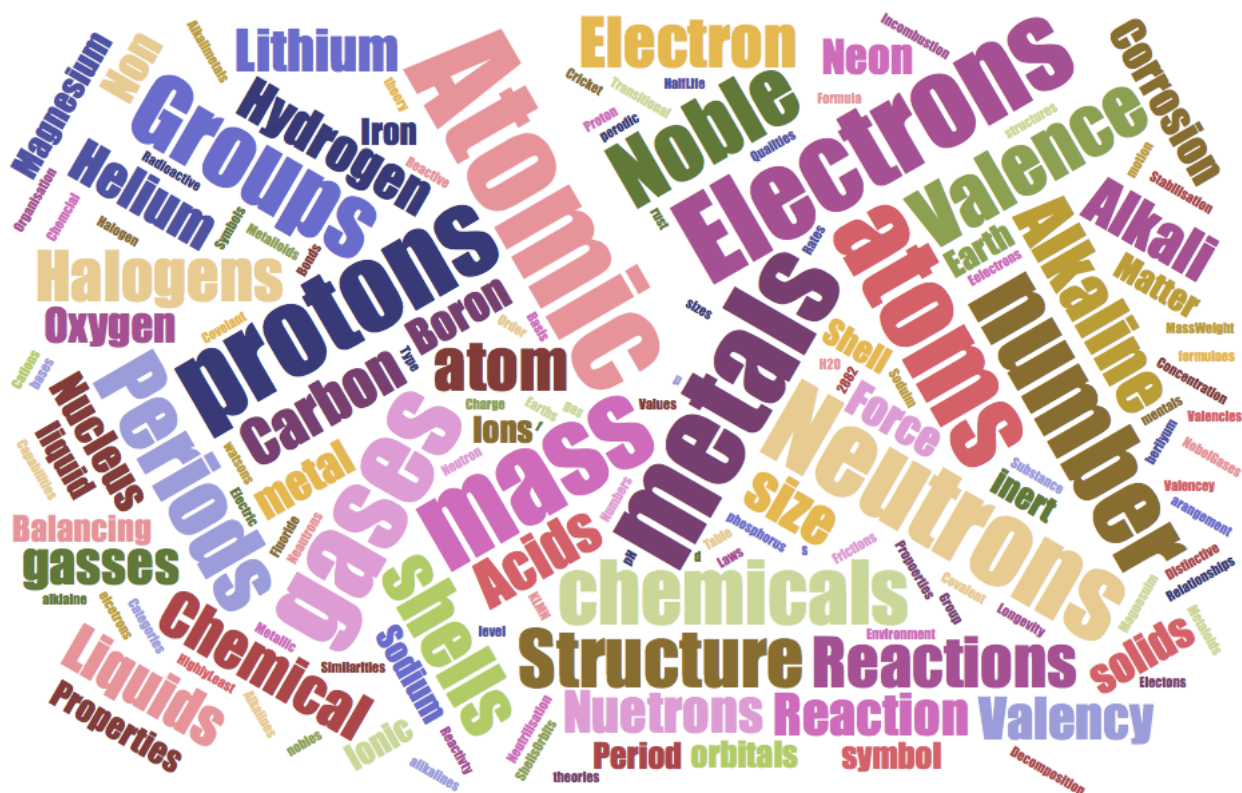


Figure 5.2. A word cloud generated from the responses from the Concept Retrieval Technique for the topic “periodic table”.

## 5.5 Design considerations for an automated Concept Retrieval Technique

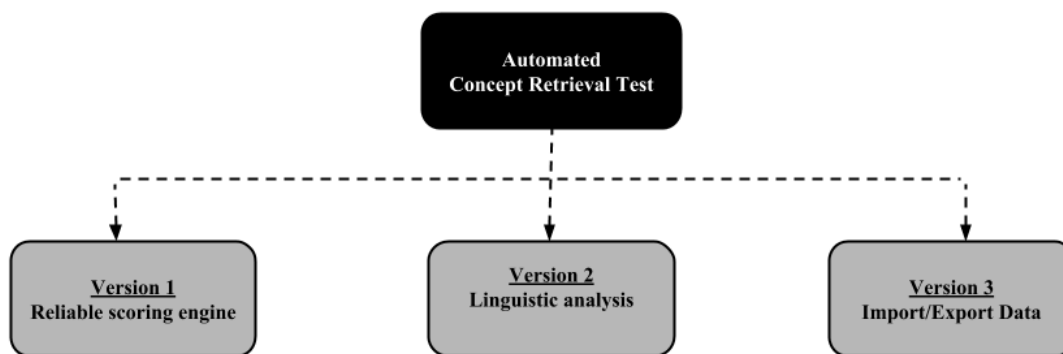
Despite the potential benefits of automated assessment, it is critical that the design of the automated Concept Retrieval Technique addresses all identified challenges and the general validity concerns with automated scoring systems. These concerns have arisen from the increased use of automated scoring systems especially in high-stake testing, which has resulted in the need for test developers to rigorously assess the validity of the inferences, based on scores produced by these systems (Clauser, Kane, & Swanson, 2002). In other words, these systems may produce reliable scores, but they may not be valid. It is therefore essential to establish sound validity evidence that supports the effectiveness of the automated assessment instrument, going beyond just the correlation scores with human raters (Cronbach, 1988; House, 1980). Considerations have been made in the design of the automated Concept Retrieval Technique, to ensure that “construct validity” issues are addressed. Ensuring that the automation process does not hinder the effectiveness of the assessment instrument.

Errors in assessment often occur when test-takers do not have a clear understanding of what is expected from them. For example, errors could occur in the automated Concept Retrieval Technique if the priming concept or subject is not clearly outlined to the test-taker. Therefore,

design considerations must be made to ensure that the implementation of the automated Concept Retrieval Technique is reliable and data collected is accurate. The first consideration is ensuring that the script that is always read out to test-takers, asking them to write down all the concepts and ideas that come to mind regarding a given topic or subject is clear within the software interface and accessible to all users. For instance, the use of label text in the software interface and audio instruction recordings can assist with consistency in the implementation of the assessment. Furthermore, test-takers are instructed to only write down keywords or bullet points and not full sentences to make scoring more straightforward and reliable. As a result, the selection of single text boxes, rather than multi-line text boxes in the software interface alleviates any issues with test-takers constructing excessive free-text responses. The second consideration is the occurrence of systematic errors made by the user and exacerbated by the strong “test-retest” reliability of machine-scoring. The use of machine-scoring eliminates the random errors that would be associated with differences among raters (or in the case of procedures that use multiple raters, differences among sets of raters). It also eliminates differences that would be associated within rating occasions. However, any errors made by the user in the administration of the software will affect all test-taker responses. Considerations will need to be made in the design of the software interface, specifically the use of help text to assist the user in minimising errors (i.e., selecting appropriate concepts in the target word list). Further considerations may involve the random sampling of users generalized across different users, environments and subjects to identify other systematic errors that occur during user testing.

### **5.6 The use of modularisation to construct the automated Concept Retrieval Technique**

Given the design considerations identified for the automated Concept Retrieval Technique the Rapid Application Development (RAD) is the most appropriate software development approach in constructing this solution. The RAD approach is used when software applications can be broken up into separate submodules for easy construction. The reasons for this selection include: project timeframe, developer expertise and access to open-source submodules such as data import/export, array searching, and visualisations and spell-checking. As a result, the software solution will be constructed by three distinct versions, utilising different submodules and features. Each version will be incrementally developed to build on the issues identified from the previous version. See Figure 5.3 for a schema of the development versions.



*Figure 5.3.* The design schema outlining the three versions of Concept Retrieval Technique development.

The primary aim of the automation process is to develop a reliable and valid scoring engine that produces consistent results with human raters for implemented Concept Retrieval Techniques. The key objectives for each version have been identified and are discussed below.

### 5.6.1 Version 1 – Reliable scoring engine

The primary aim of version 1 was to construct and test a reliable automated scoring engine that simulates human raters. This version requires a number of manual steps in the administration of the assessment and collection of test-taker response data. After test-taker response data has been collected it is imported by the user into an online database table. Then the score of each test-taker is calculated and the online database table is updated to reflect the changes. Critical in the design of the database schema and processes associated with test administration was ensuring that each test-taker response has a unique identifier. This identifier allows users to search for individual test-taker responses when the scoring has been executed. Therefore, during the importing process a unique identifier was generated based on imported student data (i.e., lastname, firstname and a random number). Test-taker responses as instructed during the implementation of the Concept Retrieval Technique were single words or sentences. Therefore, the database was designed to handle up to 400 characters for each user response and minimise test-takers writing excessive responses. Target word lists are entered by teacher and stored in an array to be retrieved when the scoring engine is activated. The scoring process commences with the first target concept and searches the first test-takers response for the occurrence of the target concept. If the scoring engine located the concept in the test-takers responses then they would be assigned one mark. This mark would be added to the test-takers total score which is stored within the test-takers database record. However, if a target concept is not located in the test-takers responses, then no mark will

be assigned to the test-taker. In both instances the next test-takers data will be checked and the process continued until there are no more test-takers in the database. This process continues by moving to the next target concept and repeating the iteration through test-taker responses. Finally, the scoring process will cease once the target word list has been exhausted and the total scores for each test-taker are available for the user in the output.

### 5.6.2 Version 2 – Linguistic analysis

The primary aim of version 2 was to address the challenges identified in the reliability of automated scoring in the Concept Retrieval Technique. Therefore, this version will provide users with the opportunity to construct a word cloud visualisation from the test-taker responses, prior to the user submitting the target word list. This assists the user to examine any consistent spelling issues or language differences (i.e., differences in American and English spelling). This version also has a number of manual steps in the test-taker response collection, however it is during the uploading of test-taker response data, that a word cloud visualisation is generated. This visualisation provides users with information on frequently used concepts within test-taker responses. This information will inform the construction of the target word list, identify possible spelling or grammar issues and provide feedback to the user on the conceptual knowledge achieved within the learning episode. For example, the target concept may be *gas*, however the user notices on the word cloud that a number of test-taker used the term *vapour*. As a result, the user may wish to score both concepts by adding them to the target word list. Furthermore, a data-cleaning submodule has been utilised in this version, focusing on identifying potentially incorrectly spelt words and flagging them for the user. This process provides the user with the opportunity to make a decision on whether the proposed new spelling of a test-taker response is valid and ensures the integrity of test-taker responses are not changed due to spelling suggestions. Finally, all feedback from the testing of version 1 has been addressed and improvements made concerning the search and scoring processes.

### 5.6.3 Version 3 – Import and export data features

The primary aim of version 3 was to provide users with the opportunity to import and export data in different formats. The import submodule will provide the opportunity for users to upload data in different formats such as excel, csv and plain text formats. Furthermore, users will have the convenience to download test-taker results after the test has been scored in similar formats. This will allow users to utilize the features of other systems such as gradebooks and

learning management systems, as all results can be manually downloaded and uploaded depending on the specified format. Finally, this version will have improved functionality addressing errors identified in the testing of version 2.

A background of a complex network graph with white and grey nodes connected by thin lines, overlaid on a light grey geometric pattern.

# 6

**Chapter**

# **Automated Concept Retrieval Technique (Version 1)**

## 6.1 Introduction

The aim of this chapter is to present and discuss the development of the automated Concept Retrieval Technique and to compare the reliability of the machine-scored measure with the reliability of human scoring. The strengths of the Concept Retrieval Technique lie in its reliability and speed of application. Therefore, the opportunity to develop an automated marking process, will provide teachers with the ability to reliably measure students learning through the access of real-time data within learning episodes. As a result, teachers can make immediate adjustments to teaching, targeted to student levels of cognition and understanding. First, a detailed background of the design process for construction of the automated Concept Retrieval Technique (Version 1) will be reported and then findings from its initial trials.

## 6.2 Design objectives

The primary aim of the automated Concept Retrieval Technique (Version 1) is to construct a reliable and valid automated scoring engine, that effectively scores test-taker assessment data based on the target word list supplied by the user. To ensure that this is feasible, the collecting and handling of assessment data inputted by test-takers will be outsourced to a suitable third-party software application. Given, the Concept Retrieval Technique has been proven to be a reliable and valid assessment instrument when manually scored, all errors that eventuate in automation will be concentrated to the machine-scoring process. Furthermore, considerations such as an appropriate data file type (e.g., csv) to be inputted into the database and ensuring the database table structure (e.g., student name, concept1, concept2) can handle the inputted data, will inform the design objectives for version 1. Therefore, the following objectives have been identified to ensure version 1 is successful in searching and scoring the correct target concepts from test-taker responses.

### 6.2.1 Objective 1 – Test-taker data collection and storage

The first objective is the effective administration of the Concept Retrieval Technique, focusing on the data collection of test-taker responses and storage in an appropriate format that can be inputted into the scoring engine. It is intended that any survey program could be used (e.g., Google Forms, Survey Monkey or Qualtrics®), provided the dataset can be saved in csv format. Because Qualtrics® was used to handle all test-taker responses during manual scoring, it will also be used to administer and collect data in the automated Concept Retrieval Technique. As a result, test-taker data collected from Qualtrics® will need to be exported in the csv format and then provisions will be made for the csv file to be inputted into the automated Concept Retrieval

Technique online database. The exported data from Qualtrics® will need to be cleaned by the user to ensure any redundant characters are removed that may cause an error in the operation of the scoring engine.

### **6.2.2 Objective 2 – Creating a dynamic online database**

The second objective will allow the users to create a dynamic database table aligned to the constraints of the Concept Retrieval Technique administration. For example, users will select the number of concepts that test-takers were required to recall and this information will be used to determine the size of the database table and name field variables. Once the table has been constructed it will allow users to input test-taker responses exported from Qualtrics®. Firstly, users will be prompted to enter the name of the unique student identifier or primary key (e.g., StudentNumber). This will ensure that users can identify test-taker scores and responses within the database. Secondly, users will be prompted to enter the number of concept fields used in the administration of the Concept Retrieval Technique (e.g., 10). In response to the concept field input the size of the table will be determined, using incrementing variables based on entered data (e.g., concept1, concept2, concept 3). Finally, a random identifier (e.g., CRTTable34) will be generated for each database table to differentiate each administration of the automated Concept Retrieval Technique. This enables all database tables to be archived and accessible for future use.

### **6.2.3 Objective 3 – Uploading CSV data files to the online database**

The third objective, is the selecting of the exported csv file and loading its contents into the online database table. The purpose of using a csv data format is that each data field is separated by a comma. Therefore, the software will parse the csv file storing single concepts or statements as unique fields. The software uses the occurrence of a comma to move to the next field. This continues for each concept field generated in the database table. Once the concept fields have been exhausted, the system will move to the next student. This process will continue until the list of students have been exhausted and feedback will be given to the user that input has been successful. An example of this process is shown below in Figure 6.1.



**CSV File**

ID1, AIDS, Ebola, Flu, Cell, Developing countries, Fringes of society, Germs, Birds, Sick, Vomit

ID2, Infection, Contagious, Dangerous, Deadly, Cures, Sickness, Ebola, Death, Cells, Traumatizing

ID3, Death, People, Animals, Spreading, Toxic, Systems, Body, Blood cells, Pain, Hospital

**Data Table**

ID1	AIDS	Ebola	Flu	Cell	Developing countries	Fringes of society	Germs	Birds	Sick	Vomit
ID2	Infection	Contagious	Dangerous	Deadly	Cures	Sickness	Ebola	Death	Cells	Tramatising
ID3	Death	People	Animals	Spreading	Toxic	Systems	Body	Blood cells	Pain	Hospital

Figure 6.1. The process of storing the Concept Retrieval Technique test-taker response csv file into the online database table.

#### 6.2.4 Objective 4 – Scoring all relevant concepts

The fourth objective is the machine-scoring process utilised in the automated Concept Retrieval Technique. The scoring engine requires the user to input the number of target concepts that have to be scored. This variable will be used to determine the size of the array that will be used to hold the target word list. The user will then enter the target word list into the system. Each concept must be saved as unique variable to be called upon independently during the scoring process. The use of Structured Query Language (SQL) will enable the software to search for the occurrence of each target concept within the test-taker response data. The process will start with the first target concept and then traverse the test-taker responses trying to find matches with the target word list. Furthermore, the SQL function will use wildcards preceding and proceeding each target concept. For instance, if the target concept is *cell* the search submodule will take into consideration any characters appearing before or after the target concept. For example, in Figure 6.2 a test-taker will receive 1 mark for recalling the concepts *cell*, *cells* or *blood cells*.

**Data Table**

ID1	AIDS	Ebola	Flu	Cell	Developing countries	Fringes of society	Germs	Birds	Sick	Vomit
ID2	Infection	Contagious	Dangerous	Deadly	Cures	Sickness	Ebola	Death	Cells	Tramatising
ID3	Death	People	Animals	Spreading	Toxic	Systems	Body	Blood cells	Pain	Hospital

Figure 6.2. Searching and identifying target concepts within the response dataset.

The search and scoring process will continue for all concepts in the target word list and for all test-takers in the response dataset. Importantly, for every occurrence of a correctly matched concept, one mark will be added to the test-takers total score. During the data input process, a total score field was created for each test-taker and the null value assigned. However, at the end of the scoring process the total score field will be updated for each test-taker. This will contain the aggregated scores from correctly retrieving concepts in the target word list.

### 6.3 Database design

Test-taker responses from each Concept Retrieval Technique need to be stored in an appropriate database to allow ease of access for the searching and scoring processes. This database needs to be online, agile and responsive to the needs of users who will require efficient and reliable scoring of the Concept Retrieval Technique. Furthermore, this database needs to be easily accessible, allowing efficient searching of test-taker responses and scoring correct responses. For version 1, this database will reside on a local server (i.e., laptop), minimising connection issues, especially when conducting alpha (e.g., developer) testing. However, in version 3 the database will be moved online, allowing a more robust beta (e.g., live environment) testing to be undertaken. The primary objective of version 1 are the decisions made from analysing test-taker responses in consideration of the target word list. Therefore, the use of a Structured Query Language (SQL) is necessary to achieve this objective, as it provides the opportunity to ask relational questions upon a dataset. The use of SQL statements will be needed to find correct responses and also update the test-takers overall score. Given the need for a scalable data management system, MAMP (Mac, Apache, MySQL and PHP) will be used as the database and installed onto a Macintosh computer to allow access to a local server environment that supports both PHP and MySQL. It is envisaged that the MAMP software will store the results of the Concept Retrieval Technique in a suitable database structure that will allow each test-taker response to be scanned and checked against the target word list to calculate an appropriate score for each test-taker.

In order for the system to be generalizable across different settings, it is important that all data collected is consistent in structure. This structure will include a unique ID field, which will be entered by the user, but in future versions this feature could be automatically generated. The ID will act as a primary key for the database, so that scores can be saved over time to measure the learning gain of test-takers. Furthermore, the ID field is of the integer data type and is needed when the score field is updated after the Concept Retrieval Technique has been scored. The name field is only used for easily identifying the test-takers and may also be used in the exporting of test-

taker results. It is character data type with a field size of 300. The test-taker response fields increment from *concept1* to *concept10*, which are of character data type with a field size of 400. The size of the test-taker response field will allow for the storage of concepts and phrases. In future versions users will have the opportunity to edit the number of test-taker response fields depending on the complexity of the concept trigger. Finally, the score is a variable that is in integer format and is the result of the calculation to determine the score for each student depending on the target word list used. See Figure 6.3 for a table structure view of the automated Concept Retrieval Technique.

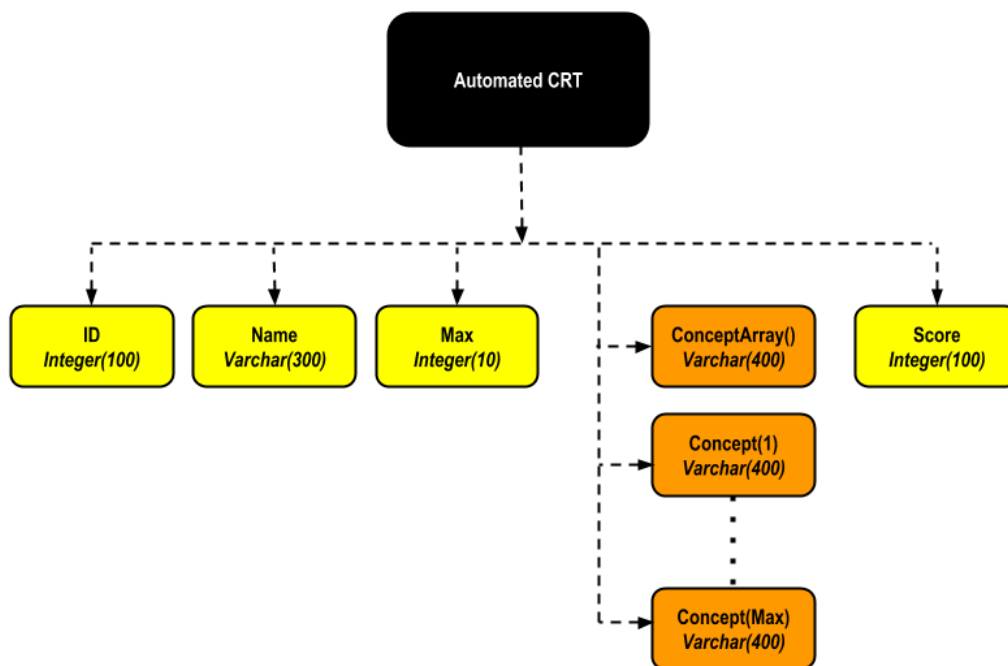
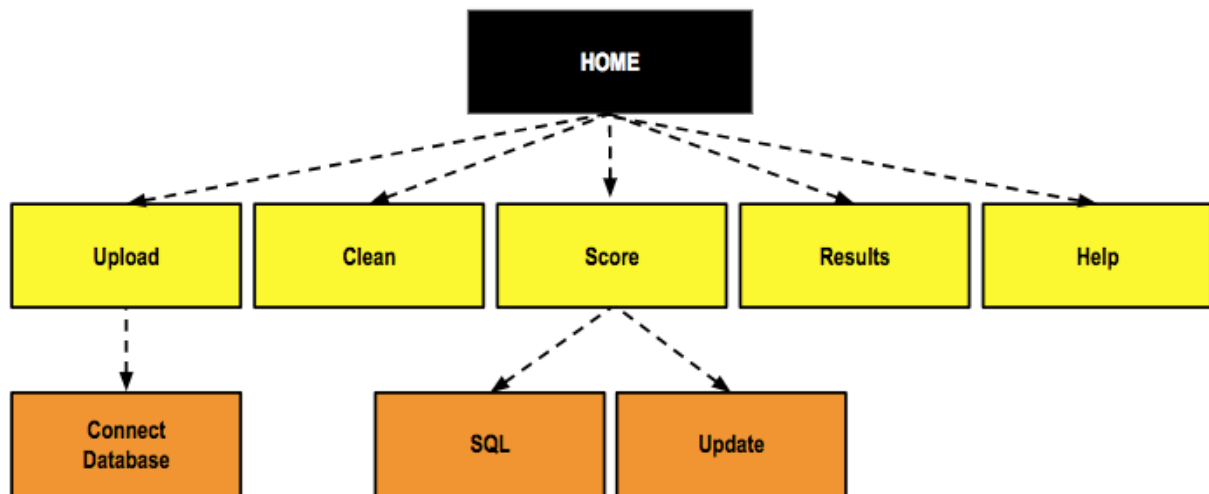


Figure 6.3. The data dictionary view of the database that holds the input from the Concept Retrieval Technique data.

## 6.4 User interface design

The design of an effective user interface focuses on usability and efficiency, ensuring errors are minimised in operating the automated Concept Retrieval Technique. Careful planning is required in developing user interfaces that ensure the user feels in control of the experience and interacts appropriately with the software. Therefore, the user interface design focuses on interaction design, visual design and information architecture. Firstly, interaction design is about ensuring that user interfaces are intuitive and interact with user, software and appropriate hardware. It is also important that developers have a clear understanding of the page relationships and utilise features such as a navigation menu to support user interaction with the software. Figure 6.4 shows a storyboard that highlights the navigation flow in operating the automated Concept

Retrieval Technique. As a result, a navigation menu will be used, which allows the user to access all features of the software and connects automatically to the database. For instance, selecting upload from the menu will connect to the database allowing the user to upload test-taker responses. The consistent appearance and location of the navigation menu is critical to ensure the operations can be handled fast and efficiently minimising user confusion or error.



*Figure 6.4.* Storyboard detailing the navigation of modules in the automated Concept Retrieval Technique.

Secondly, visual design is concerned with the aesthetics and selection of appropriate interface elements (i.e., buttons, text boxes or list boxes). The principles for creating an effective user interface include clear and appropriate help text, consistent page layout, appropriate use of icons/graphics and colour scheme selection. The construction of individual interface wireframes are beneficial to evaluate the effectiveness of proposed interfaces according to effective screen design principles. Figure 6.5 shows the interface wireframes for the automated Concept Retrieval Technique. It is important that help text is clear, simple and consistent. For example, in the create database interface users are given instructions such as “please select the number of concepts to be scored”. Following the instruction, more detail is provided to the user “this is how many concepts you have asked students to retrieve”, which is emphasized by the use of bold text and helps the user understand how the input will be used by the software. The use of a consistent page layout is important and helps users to feel empowered, enhancing the overall usability of the software. Furthermore, the use of cascading style sheets (CSS) will ensure that all colour schemes, fonts and sizes are consistent for each page in the program. The use of the navigation menu at the top of

each page will ensure that navigation is clear and not cumbersome. Finally, the graphics and colour scheme selected will utilise a clear and coherent functionality using an appropriate white space to promote a simple easy to use, yet effective software application, which is an expectation of manual Concept Retrieval Technique.

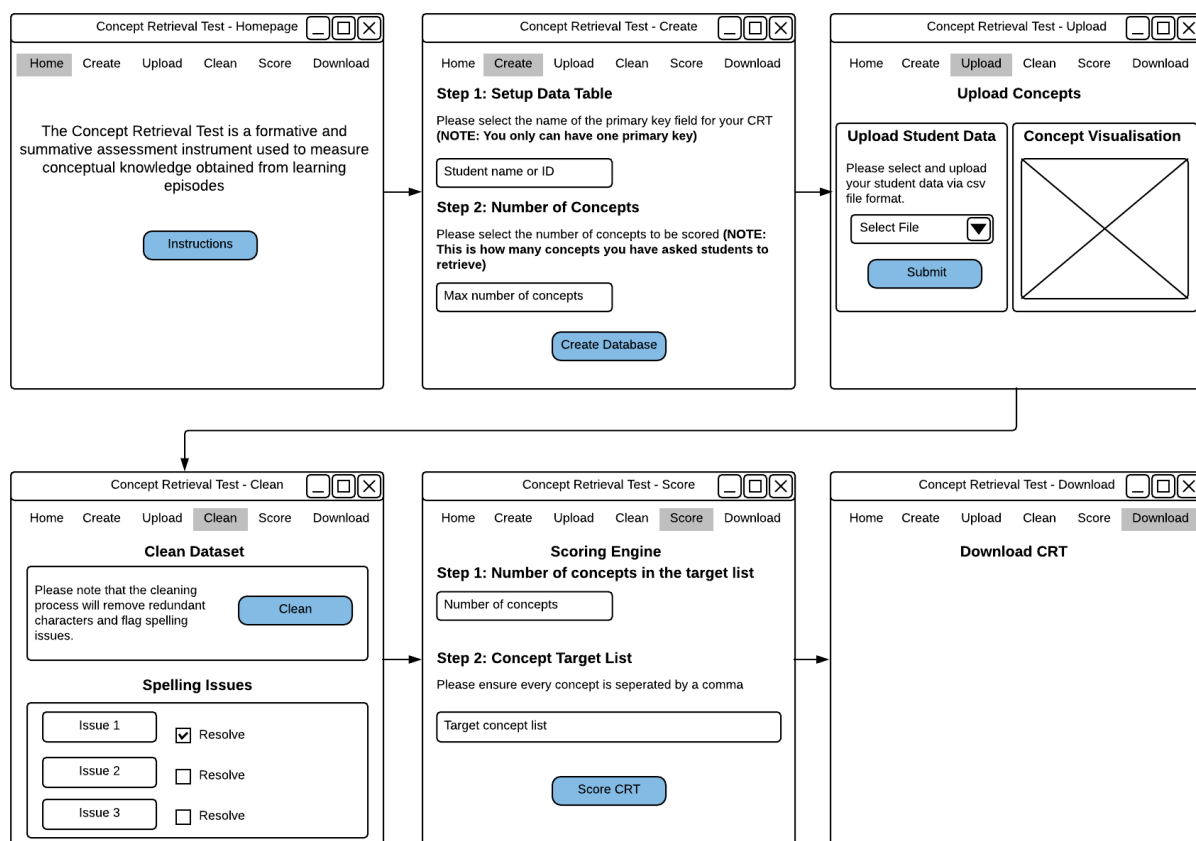


Figure 6.5. Interface design wireframe for the automated Concept Retrieval Technique.

Finally, an understanding of the information architecture is required to consider how the interface connects to external systems. This includes the ability to connect to a server and run SQL processes. The benefit of using a MySQL database is that php can be used to program the functions that are executed upon the database, from within the interface. However, considerations such as browser configuration, use of mobile devices and any third-party software (i.e., the use of flash for the visualisation) need to be examined to determine any compatibility issues that may arise.

The final aspects of interface design involve testing the usability of the created interface. Usability testing is undertaken by the developer who is familiar with the interface and identified problems are documented and feedback is used to improve the overall design of the interface. Often usability testing is not the only testing of interfaces. Using field trials and fine-tuning using a much larger group of test-takers, that more closely resembles the final product are used to uncover issues with the operation of the interface (Fulcher, 2003).

## 6.5 System modelling

The purpose of system modelling is to represent the different software submodules and related components of the entire system, including access to third-party software applications. Modelling tools are utilised to provide extensive information to developers to assist in the construction of the software solution. The most common representations include system flowcharts that provide information about all internal and external elements within the system. These flowcharts are also used to represent individual submodules and to help gain a deeper understanding of how each version objective is achieved.

### 6.5.1 The automated Concept Retrieval Technique (Main Module)

The main module (i.e., the highest level in the hierarchy) is the driving module that the user will have the initial interaction with. Figure 6.6 represents a system flowchart for the main module of the automated Concept Retrieval Technique. This representation shows the initial process (e.g., Connect to phpMyAdmin) that makes a connection to the storage facility (e.g., CRT SQL database). This connection enables data to be stored (e.g., test-taker responses) and updated (e.g., Total score) depending on the user's navigation. The user is responsible for the operation of the system, specifically selecting submodules to be activated (e.g., Create database). All submodules related to version 1 have been isolated in the colour blue and subsequent flowcharts constructed for each submodule to provide further explanation.

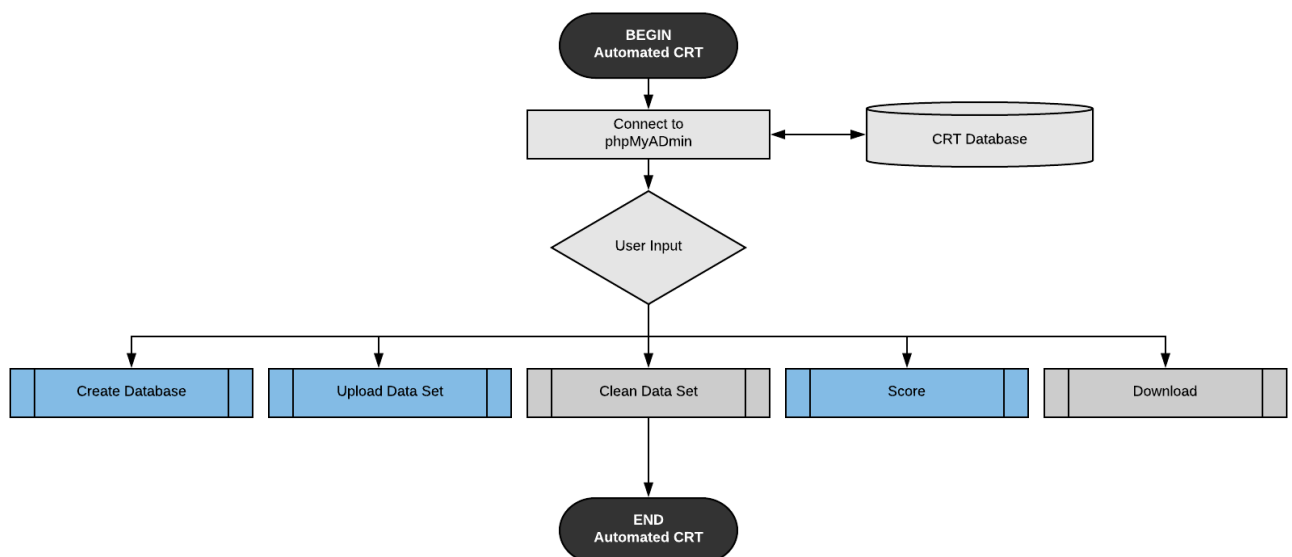


Figure 6.6. System flowchart of the automated Concept Retrieval Technique with the key submodules in version 1 highlighted.

### 6.5.2 Create database (Submodule)

The purpose of the create database submodule is to address design objective 2. Figure 6.7 represents the system flowchart for the create database submodule, with the first process ensuring the connection to the CRT MySQL Database is active. In order for the table to be created, two variables are required to be inputted by the user. The first is the name of the primary key (student identifier) for the database table. The purpose of the primary key is to allow the user to differentiate the responses and scores from test-takers. Initially, this will be important in judging the validity of the automation, by highlighting any differences between human and machine-scoring of individual test-takers. Furthermore, this will allow users to identify test-takers who may require intervention in learning based on their responses. The second variable is the number of concepts collected from the test-taker response dataset. Users have autonomy over the number of concepts recalled in the administration of the Concept Retrieval Technique, ensuring that it represents the amount of conceptual knowledge addressed in the learning experience. Therefore, the number inputted by the user when creating the database table will be used to construct the number of columns and unique field names such as concept1, concept2, concept3. Each field generated to contain test-taker responses will be assigned a string data type of 400 characters and the total score field will be assigned a numerical data type of three characters. Another submodule is called that creates the database table in the online database. Finally, an output is determined based on the status of the table creation. In both instances the user is notified.

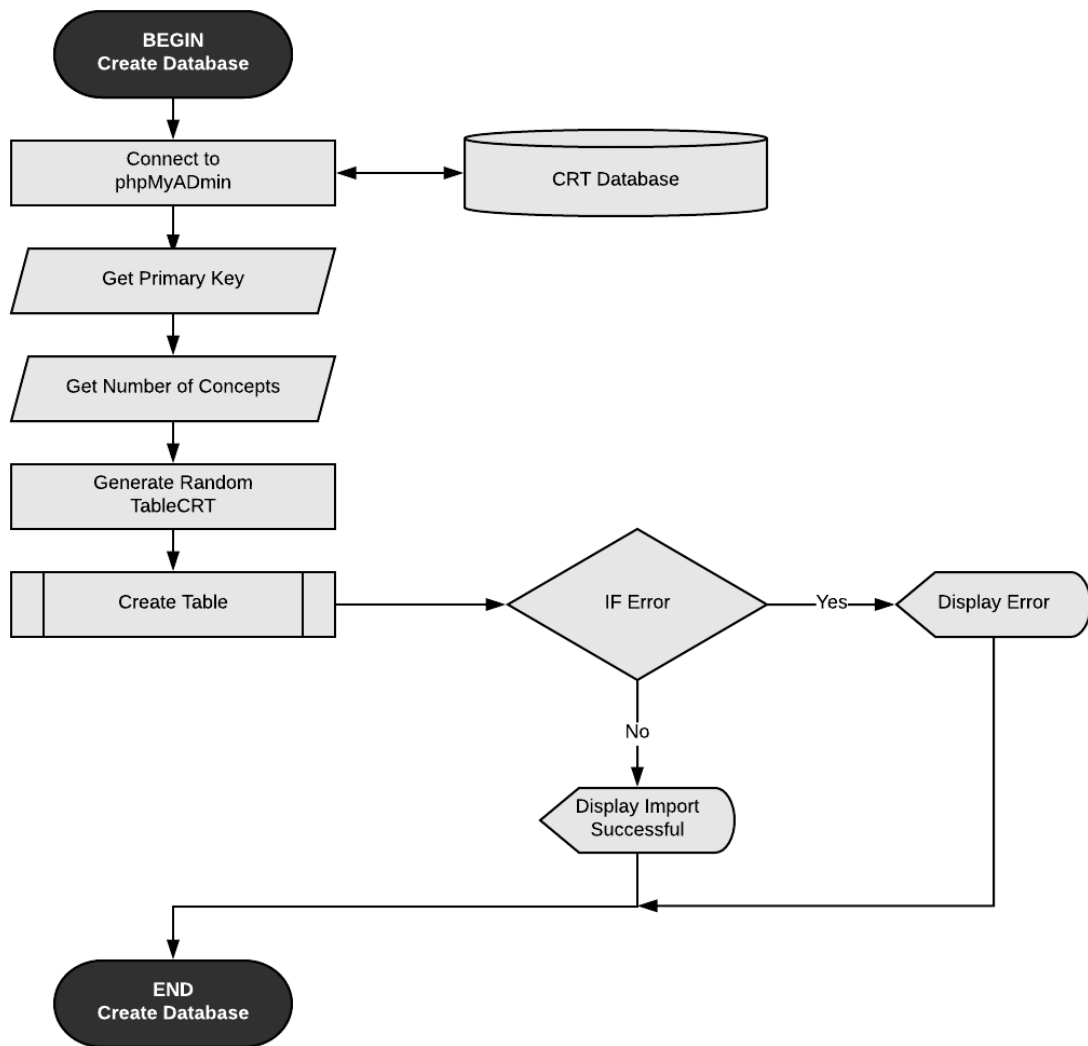


Figure 6.7. System flowchart for the create database submodule.

### 6.5.3 Upload data set (Submodule)

The purpose of the upload data set submodule is to address design objective 3. Figure 6.8 represents the system flowchart for the upload data set submodule. Similarly, to other submodules the status of the database connection is verified as the first process. The user is then prompted to select a file that utilises the csv format. Once the file has been verified then the input data submodule is called to store all test-taker responses in the identified database table. It is critical that the format of the inputted file matches the format of the table. Otherwise, runtime errors will occur frequently. A common problem is the user mistaking the number of responses in the database table creation. Therefore, any issues identified with the upload process are provided as feedback to the user through the user interface.



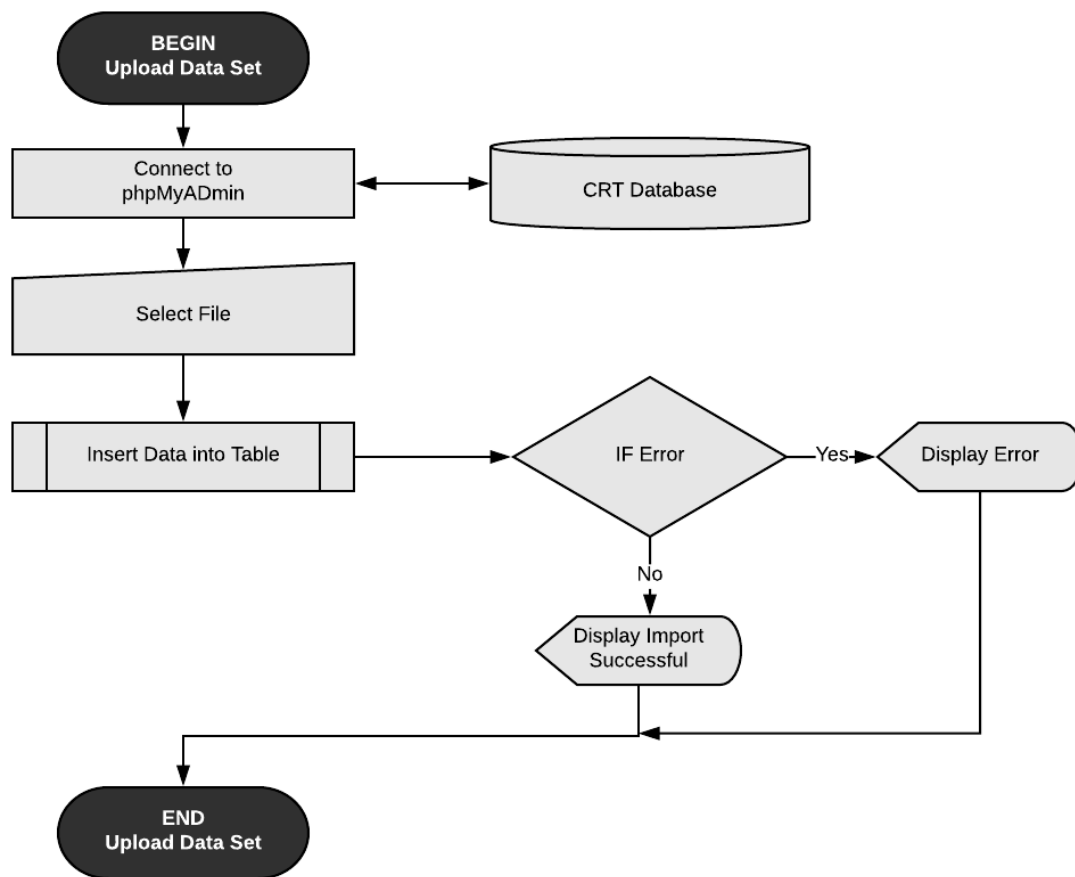


Figure 6.8. System flowchart for the upload data set submodule.

#### 6.5.4 Scoring engine (Submodule)

The purpose of the scoring engine submodule is to address design objective 4. Figure 6.9 represents the system for the scoring engine submodule. In this submodule the user is required to submit two inputs. The first input identifies the number of concepts in the target word list. This variable is used to inform the number of iterations that are undertaken when comparing target concepts to test-taker responses. For instance, if six is entered then the software recognises that there will be six concepts that need to be checked in every test-takers response list. The second input is the target word list. This information is inputted as a single line of text, with each concept separated by a comma. The target word list will be stored in a temporary single dimensional array. The score concepts submodule is called and will check the occurrence of each target concept in test-taker responses for every test-taker. If a concept match is found then a mark is added to the test-takers total score. Finally, the total score is rewritten back to the online database and if any errors occur the user will be notified.

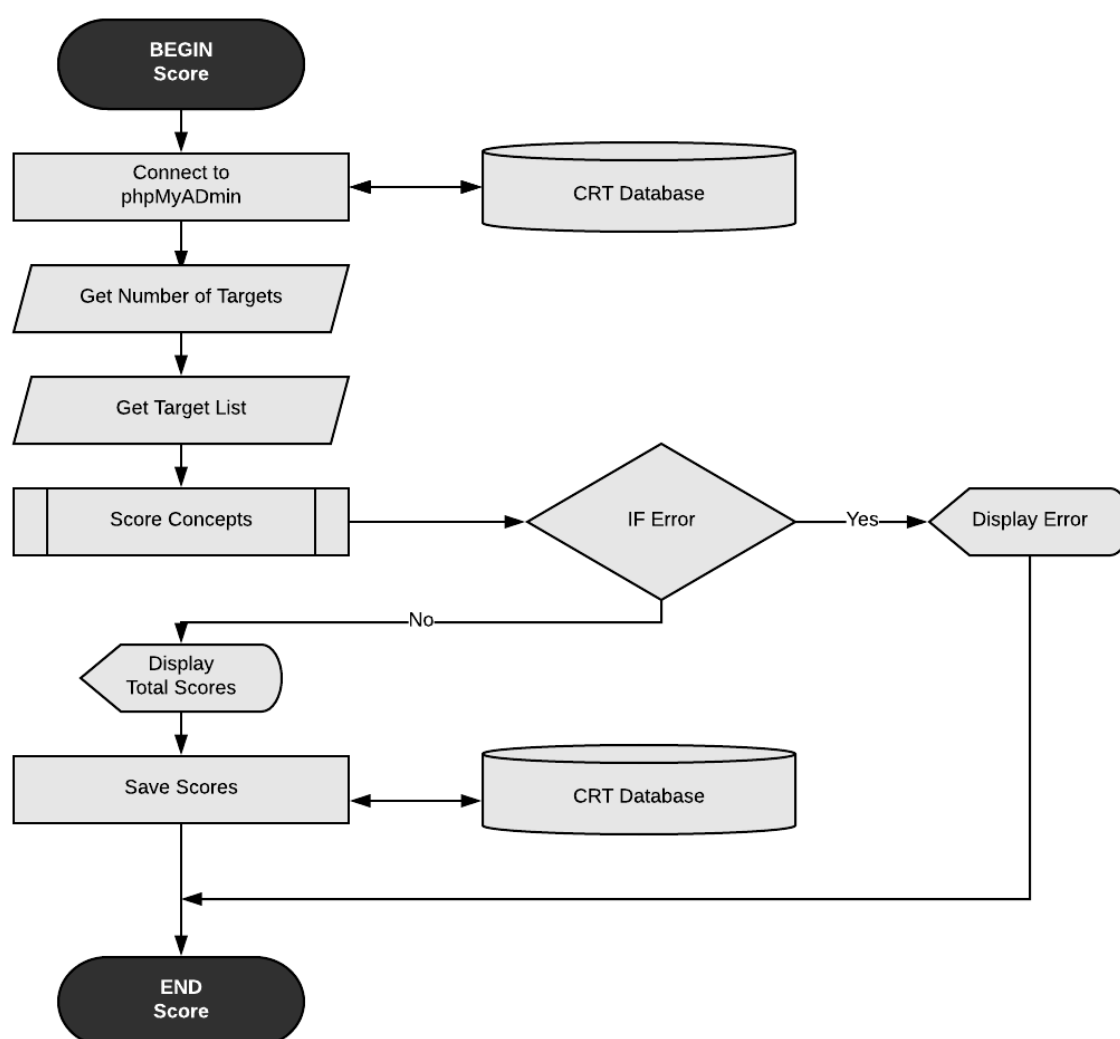


Figure 6.9. System flowchart for the score dataset submodule.

## 6.6 Module construction

Using the system modelling tools, each submodule, is created using php, javascript and html to achieve the stated objectives for version 1. Algorithms are an important tool to help developers understand the inputs, outputs, processes and decisions within the submodule. Therefore, algorithms have been provided in this thesis to help understand how the objectives have been achieved in the creation of the automated Concept Retrieval Technique.

### 6.6.1 Database connection (Submodule)

In order to allow test-taker responses to be uploaded into the database for scoring to be undertaken, a connection to the phpMyAdmin database needs to be established. The connection is established via php and this code is called when the user processes a file to be uploaded. Javascript

is commonly used to provide file processing capabilities in an online environment and will be used for the file selection and uploading. Figure 6.10 provides an algorithm representation of this submodule, showing the connection to the phpMyAdmin database. Once the connection submodule has been called, variables such as database name, username and password are required to establish the connection. If they are incorrect an error message is logged. However, in both instances (i.e., connection established or no connection) feedback will be provided to user. Figure 6.11 provides a visual representation of the interface with the javascript file processing screen elements. In this example a successful connection with a database has been established.

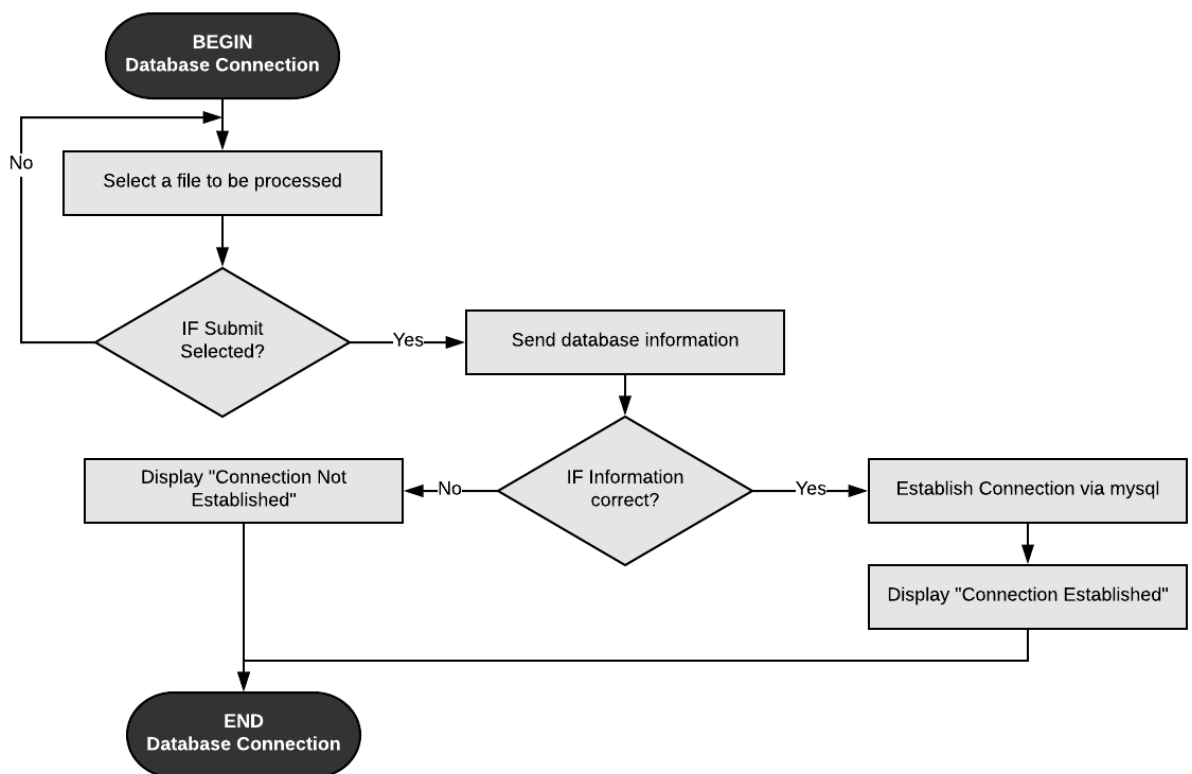


Figure 6.10. Algorithm representation for the database connection submodule.

Connection established Database selected  
Choose your file:

No file chosen

Figure 6.11. Interface for database connection submodule with feedback to the user regarding connection.

### 6.6.2 Upload dataset (Submodule)

This submodule will allow the user to select the csv file of test-taker responses downloaded from Qualtrics® and upload its contents to an online database table. During the upload process connections are made to the online database table to ensure that the field data in the csv file (i.e., separated by a comma) match the structure of the MySQL database. If this is incorrect, it could have a significant effect on the validity of the machine-scoring. For example, if test-taker responses were mixed up the final scores may not be reflective of student conceptual knowledge. The csv file will contain information such as test-taker id and responses separated by a comma. See Figure 6.12 for a screenshot of the data collection interface on a mobile device and a sample of the dataset in csv format.

1	David	Silicon-based Photons	Energy effici	Solar battery	Sustainability	Parcels	Greenhouse	Printable PV
2	Yaghi	Silicon based Photons	Energy effici	Solar battery	Sustainability	Parcel	Green house	Printable PV
3	Thomas	Convert in el if you connect	Less than 20'	They contain	The energy p	They have no	They do not	An Inverter is
4	Patrick	Photovoltaic Converts the	The absorpti	Electromagn	Series and p	DC voltage	Power boxes	Size of solar
5	Holsinger	Energy Photons	Photovoltaic	Silicon-based	Cells	Solar	Current	Electron
6	Miranda	Photons Photovoltaic	Silicon-based	Cell	Electron	Current	Panel of Grid	Energy
7	Joseph Thorn	Series Circuit Positioning	Area	Expenses	Weather	Voltage	No greenhou	50 years
8	Kanbour	Parcels Silicon based	Photons	Solar radiatio	Energy effici	Circuits	Solar battery	Inverter
9	Duff	Photovoltaic Volts	Direct Curre	Series/ paral	Photons	Efficiency	Inverter	Cells
10	Silva	Conversion Energy	Radiation	Volts	Light	Solar farms	Solar Panels	Positioning
11	Revollar	Photons Conversion	Area	Volts	Light	Energy	Solar Farms	Output
12	Alfonso	Photons Solar	Voltage	Sunlight	Electrical	Efficiency	Sustainability	Heat
13	han	Photovoltaic Energy	Silicon	Panels	Kilowatts	DC	Photons	Light
14	quharson	Silicon Sunlight	Photons	Series	Parallel	Efficiency	Voltage	Inverter
15	elieka	Solar Cells	Light	Energy	Electricity	Power	Efficient	Photons
16	Patel	Reproduction	Efficiency	Cells	Advantages/	Type of solar	Photons	Wheather
17	Quisumbing	Radiation	Transformati	Voltage	Efficiency	Wires	Photovoltaic	Thermal
18	Francis	Photons Heat energy	Light energy	Silicon based	Electrical ene	Reusable ene	Voltage	Series Circuit
19	Raffoul	Photons Efficiency	Varied inter	Heat energy	Electrical ene	Silicon based	Series circuit	Parallel circu
20	Valentino	Angle Sunlight	Series	Silicon based	Voltage	20% efficien	Light energy	Inverter
21	Amurao	Surface area Cost	Light	Direction	Income	Voltage	Charge	Variables
22	Jiang	Energy Volts	Photovoltaic	Series/Parall	Photons	Heat	Energy effici	Cells
23	Newman	Energy Trans Photons	Photovoltaic	Solar Cells	Efficiency	Power Sourc	Circuit	Chemical En
24	Khoury	Photovoltaic Angle	Connection	Area	Shadow	Mains	Clouds	Fusebox
25	Matta	Sun Solar power	Direction of	Voltage	Power	Energy	Heat	Energy
26	Miralles	Solar energy Heat energy	Circuits	Voltage	Weather	Series and P	Engineering	Solar power
27	Raffoul	Silicon Based Series	Parallel	Voltage	Heat	Electrical	Energy	Photons
28	Saba	North Silicon	Series	Parallel	Voltage	Heat	Electrical	Photons

Figure 6.12. The Concept Retrieval Technique data collection process.

The process to ensure that the uploaded data is valid, requires the storage of characters into a string and then storing that data in a field within the database at the occurrence of a comma for every student. Figure 6.13 provides the algorithm representation for this submodule with the first process calling the data collection submodule to establish a connection. The user is required to select the relevant csv file to be upload and this is checked to determine if data is available. If there is data in the file each character scanned using a pre-test repetition statement and adding the characters to a string until it recognises a comma value (,). This will flag a field change (i.e., username to Concept1) and the string will be saved to the database table. The process is repeated

until there are no more fields available. This causes the input to move to the next student with the process repeating again only if there is data available to input.

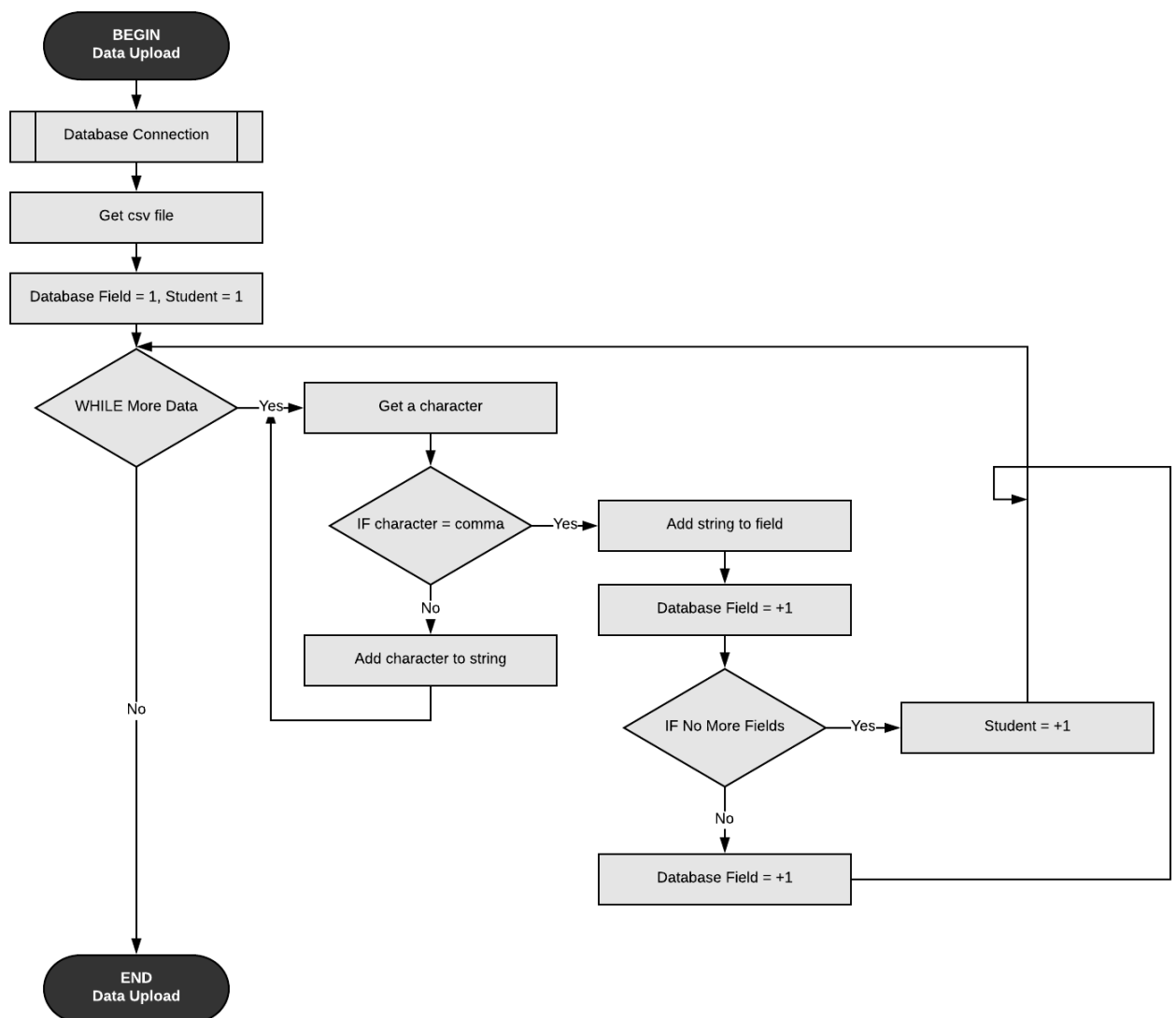


Figure 6.13. Algorithm representation for the data upload submodule.

The final iteration of the automatic upload submodule utilised the cascading style sheets and more advanced data input functionality. See Figure 6.14 for Interface for the upload dataset submodule. The execution of the submit button loads the selected csv file into temporary memory. It is then scanned to check that the format is correct and the concepts are loaded into variables including the username, which is used as a primary key to ensure data searching and sorting capabilities. Then using SQL functionality, each string variable is imported into the MySQL database, allowing other submodules within the program to access and manipulate it. Finally, feedback is given to the user regarding the success of the import feature.

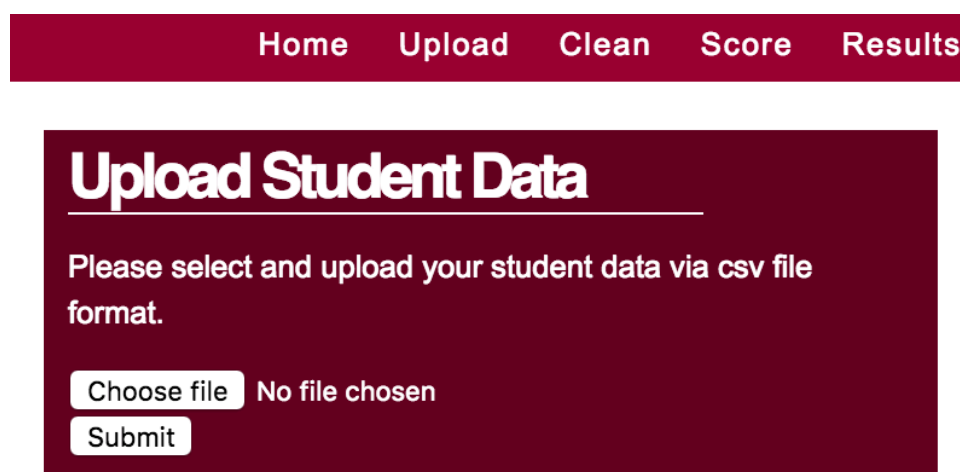


Figure 6.14. Interface for the upload data submodule with CSS styles and navigation menu.

### 6.6.3 Saving the target word list (Submodule)

Once the database has been populated with the test-taker responses, the next step in the automation process is to enable users to input the target word list. Users are presented with input text boxes to enter target concepts and they are stored as single strings within a temporary one-dimensional array. A search submodule will then be used to scan the MySQL database of test-taker responses to identify matches to the target word list criteria (i.e., *concept1 like targetconcept1*). It is anticipated that spelling errors will not occur in the inputting of the target word list, however future evolutions of this program could include a spell-checker for target word list concepts. Figure 6.15 provides a representation of the algorithm for this submodule starting with user entering the maximum number of target concepts. This variable is the flag for the loop that will iterate through each target concept and store them in a temporary data structure. Each time a concept is inputted the variable “target” is incremented. When this variable is no longer equal to the maximum number of target concepts variable the loop will end. Figure 6.16 shows the interface for this submodule. Input text boxes have been utilised to capture the inputs of the user for the scoring submodule. Instruction labels were also added to ensure that the user has an adequate understanding of his or her role within the automation process. Specifically, the scoring submodule will use this information to generate a score for each student to determine his or her conceptual knowledge.

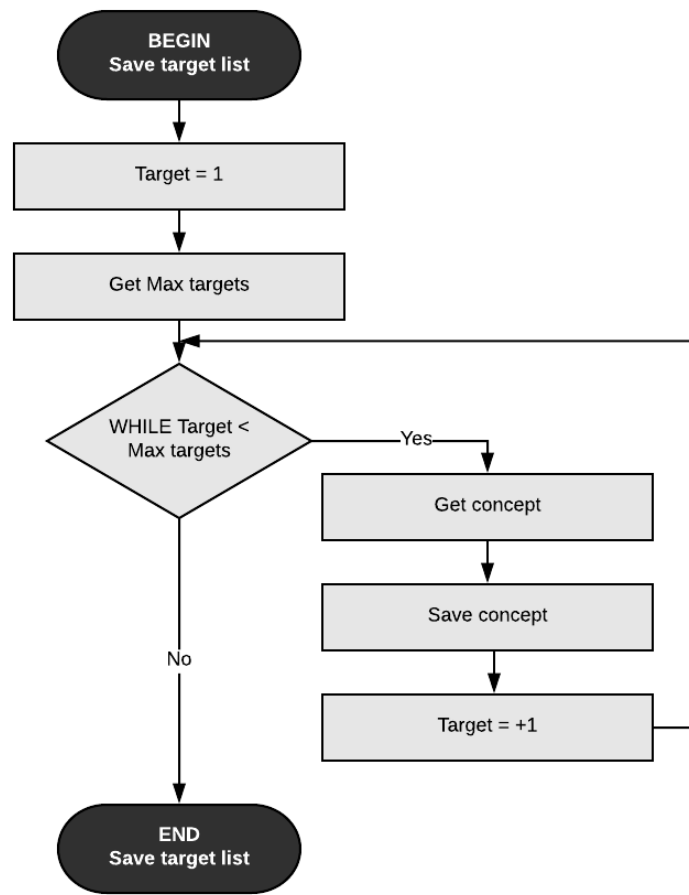


Figure 6.15. Algorithm representation for the saving target word list submodule.

## Concept Recall Test Online

### Score the CRT

Please enter the concepts that you wish to score in the text boxes below.

<input type="text" value="Enter concept here...."/>	<input type="text" value="Enter concept here...."/>
<input type="text" value="Enter concept here...."/>	<input type="text" value="Enter concept here...."/>
<input type="text" value="Enter concept here...."/>	<input type="text" value="Enter concept here...."/>
<input type="text" value="Enter concept here...."/>	<input type="text" value="Enter concept here...."/>
<input type="text" value="Enter concept here...."/>	<input type="text" value="Enter concept here...."/>

Figure 6.16. Interface for the saving target word list submodule.

#### 6.6.4 Searching for a target concept (Submodule)

The searching submodule is the first part of the scoring engine. It is a critical element of the program to search test-taker responses for the occurrence of target concepts. It is imperative for the program to transverse the database table of test-taker responses to match the concepts in the target word list and score them accordingly. Therefore, it was anticipated that if it can find one concept correctly then the code could be scaled appropriately to include multiple concepts. The theory for this submodule is that every concept submitted by a test-taker will be checked against every target concept. Figure 6.17 shows the search process by starting with the first concept *Atom* from test-taker 1. This concept is compared against the first target concept *Atom*. If a match occurs a message will be outputted to the user such as “target found” and the score will be incremented. This process continues if a concept is not found by comparing all target word list concepts against the first response by test-taker 1, until the list of target concepts has been exhausted. Figure 6.18 shows the iteration of the loop by moving to the second concept *formula* from test-taker 1. The comparison will start back at the beginning of the target word list and check the concept against every concept in the target word list. This process will continue until every concept from test-taker 1 has been exhausted. Finally, this will also continue for every test-taker in the database. Importantly, the time taken to perform these comparisons is minimal in a machine-scoring environment, highlighting the efficiency of automated assessment.

		Target Concept List									
		Atom	Proton	Mass	..	..	..	..	..		
										Score	
Test Taker 1	Atom	Formula	Period	Proton	Mass	Table	..	..		1	
Test Taker 2	Shell	Atom	Mass	Atom	Element	Reaction	..	..		0	
Test Taker 3	Shells	Promton	Shells	Atom	Atomic	Form	..	..		0	
Test Taker 4	Protons	Protos	Mass	Periodic	Atomic	Period	..	..		0	

Figure 6.17. The first iteration of the scoring submodule, where the first concept of test-taker 1 is checked against all target concepts.



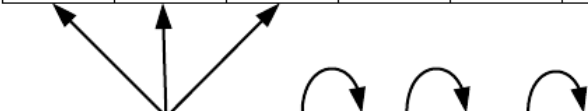
Target Concept List									
	Atom	Proton	Mass	..	..	..	..	..	
									
Test Taker 1	Atom	Formula	Period	Proton	Mass	Table	..	..	3
Test Taker 2	Shell	Atom	Mass	Atom	Element	Reaction	..	..	0
Test Taker 3	Shells	Promton	Shells	Atom	Atomic	Form	..	..	0
Test Taker 4	Protons	Protos	Mass	Periodic	Atomic	Period	..	..	0

Figure 6.18. The second iteration of the scoring submodule, where the second concept is checked against all the target word list concepts.

The first iteration of this submodule is the ability to search through the database and find the occurrence of one target concept. Figure 6.19 shows the algorithm representation of searching for a concept in the target word list. Similar to previous submodules the first process is calling on the database connection submodule. This establishes a connection with *tableCRT* in the MySQL database allowing relational questions to be asked via SQL commands (i.e., `SELECT * FROM tablecrt WHERE concept1 LIKE Target1`). The next step requires the initialisation of the size of the target word list (e.g., `MaxTarget`) and the current array pointer for the target word list (e.g., `TNum`). The first decision requires the current target concept to be less than the maximum size of the target word list, before the loop is entered. If this is true then the comparison will be executed. Following this comparison, a decision will be made informing the user if the target concept is found. Resulting in the incrementing of the `TNum` variable and changing the array pointer to the next target concept. Finally, this process will continue until there are no more target concepts within the target word list to compare.

Specific detail concerning the comparison decisions are not represented in this algorithm. However, these decisions will need to be programmed into the scoring submodule. During the execution of this submodule it is anticipated that test-taker responses will not always be an exact match. The ability to derive a target concept match within test-taker responses is made by human raters and needs to be replicated by the machine-scoring process to ensure similar levels of validity. Therefore, wildcard characters will be utilised and placed at either side of the target concept to enable the target concept to be extracted from the response. For example, if the target concept is *cell* then following concepts would be scored correctly, *cells*, *blood cells* and *plant/animal cell*. It is anticipated that the use of wildcards will also minimise issues with the use of plurals and rectify

some spelling issues. Figure 6.20 shows the interface for the search concept submodule, it was trialled searching for the test-takers first or last name and provided users feedback if the target was found. The programming used within this submodule will be used by the search submodule within the automated Concept Retrieval Technique.

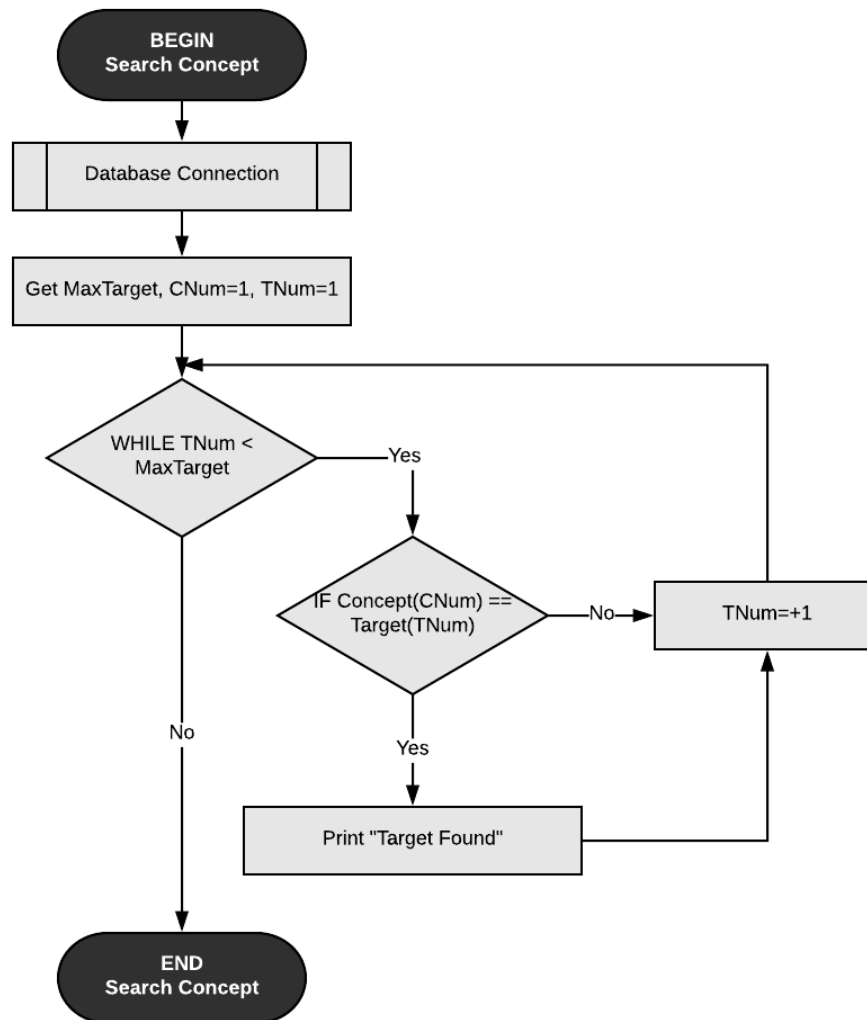


Figure 6.19. Algorithm representation for the search target concept submodule.

Connection established Database selected

## Search Contacts Details

You may search either by first or last name

Figure 6.20. Interface for the search target concept submodule.

Although this submodule was effective in identifying if a target concept was found, there were two issues identified that limited its effectiveness. The first issue highlighted is the lack of feedback provided to the user during the execution of the submodule. Specifically, when a target concept is not found within the test-taker responses, no feedback is outputted to the user and it assumed that the search was successful. Furthermore, the output was only limited to identifying if a target concept was found not the frequency of that found concept within the test-taker responses. In future versions it is important that appropriate feedback is provided to the user regarding the outcome of all actions. Specifically, in the next version a variable will be added to the loop that increments every time a target concept is found. This variable will also provide feedback to the user regarding how many target concepts were found within test-taker responses. Furthermore Figure 6.21 shows the user feedback when a target concept is not found. This provides feedback to the user that the target concept was not found, not an error in the operation of the software. The second issue resides in the looping through the target word list. Even if a match is identified the program still loops through the test-taker responses. This may cause issues with validity as test-takers may receive multiple marks for the correct concept appearing multiple times within their responses (common with forced recall).



Figure 6.21. Interface for the search target concept submodule.

### 6.6.5 Scoring engine (Submodule)

For the scoring engine submodule to be effective changes needed to be made to iterate through each test-taker response for every test-taker in the dataset. Figure 6.22 provides a representation of the algorithm for this submodule. Similar to previous submodules the database is connected. Within the interface there are ten search fields established to contain the target word list items. However, the next version of the program will not be constrained to ten target concepts and users will have the option to vary the size of the target word list. The loop will start with the first response of the first test-taker and will iterate through the concepts in the target word list to identify concept matches. The comparisons made between test-taker responses and target word list concepts are made using the following code example “\$resultSet = \$mysqli->query(“SELECT \* FROM tablecrt WHERE concept1 LIKE ‘%\$search1%’ OR concept2 LIKE ‘%\$search1%’);”

which provides a snapshot of two concepts in the target word list. The result of a concept match will cause the score variable for that test-taker to increase by one mark. It is important that the row is correct so that it updates the score for that test-taker. The process continues for each concept entered by a test-taker until there are no more fields to check for concepts and then the total score is saved and the row is incremented to take on the next test-taker in the database. The database table is updated using the following code “\$sql = \$mysqli->query("UPDATE tablecrt SET score = \$score WHERE CRTID = \$CRTID");”, which updates the score variable in the database score field. Finally, Figure 6.23 shows the interface design for the scoring submodule, utilising text boxes to minimise test-takers submitting complex sentences.

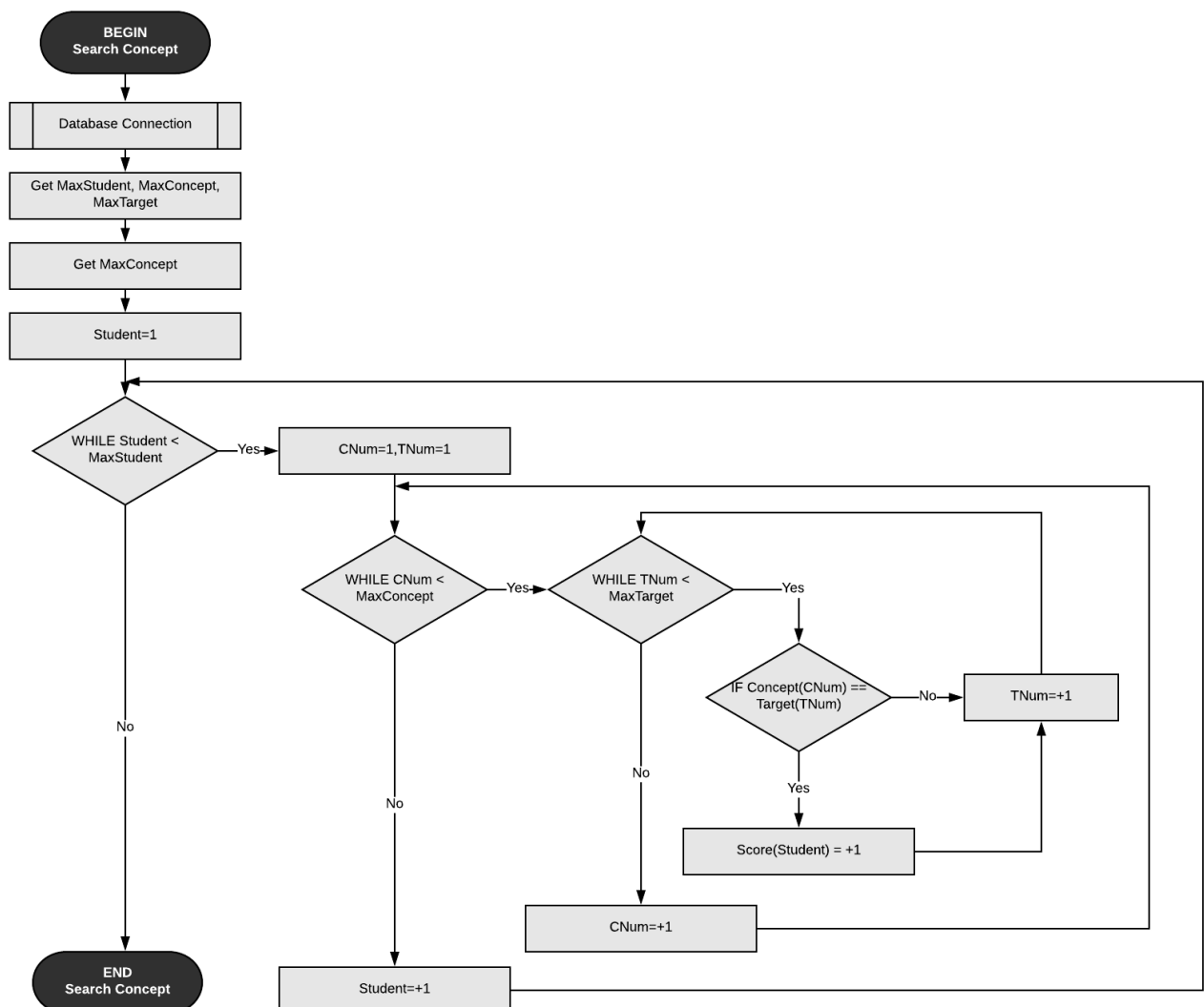


Figure 6.22. Algorithm representation for the search and score submodule.

**Score the CRT**

Please enter the concepts that you wish to score in the text boxes below.

Enter concept here.... Enter concept here....

Enter concept here.... Enter concept here....

Enter concept here.... Enter concept here....

Enter concept here.... Enter concept here....

Enter concept here.... Enter concept here....

Score CRT

Figure 6.23. Interface for the scoring submodule.

### 6.7 Study 6 - Reliability of the automated Concept Retrieval Technique (Version 1)

In the current study, we explored the feasibility of machine-scoring applied to the Concept Retrieval Technique. The primary purpose of this study was to determine the overall reliability of the machine-scoring in comparison to human scoring. We used Cohen's kappa as the statistical measure to examine the correlation that indicates agreement and correspondence between machine to human scores. Furthermore, this measure provided a determination of the inter-rater reliability and overall utility of the automated assessment measure. The test-taker concepts from a previously scored Concept Retrieval Technique were imported into the automated Concept Retrieval Technique (Version 1) and using the same target concept a total score was generated for each test-taker. The results of two independent raters were used to generate an agreed score, which was analysed to determine the consistency of the inter-rater agreement between both machine and human scores

## Method

### Participants

Fifty-four secondary school science students participated in the study from an all-male secondary school in Australia. Their average age was 15 years ( $SD = .52$ ). In test two, forty-five secondary school science students participated from the same school. Their average age was 13 years ( $SD = .32$ ). Participation was voluntary and students were not compensated for their participation.

### Materials

A Concept Retrieval Technique was developed for the topic of the “periodic table”. A subject-matter expert worked together with the researchers to generate the target word list containing all admissible concepts for the topic. All aspects of this study are detailed previously in study 4 of this thesis. The automated Concept Retrieval Technique (Version 1) was used to machine-score the test-taker responses to be used in this study.

### Procedure

The administration of the Concept Retrieval Technique included an instruction, requiring students to write down all concepts or ideas about a topic using only keywords or bullet points, with the directive to avoid writing in full sentences. Following the administration of the Concept Retrieval Technique, two raters independently marked the responses with the help of the target word list. One mark was awarded for each correctly identified concept and a final score for each student was thereby generated. The inter-rater agreement generated from the two raters for the Concept Retrieval Technique was  $\kappa = .85$ . The raters then resolved any inconsistencies in their scores to generate a final agreed score. This score was used as a measure of inter-rater reliability against the scores generated from the automated Concept Retrieval Technique (Version 1). The operation of the automated Concept Retrieval Technique (Version 1) required the same list of student concepts to be imported into the online database. Furthermore, the same target word list used by human raters was imported into the program and machine-scoring executed. Finally, the total scores of the machine-scoring process were downloaded in csv format and added to SPSS to be analysed. Comparisons were made between both methods of scoring.

## Analysis

Inter-rater agreement was established by computing the Cohen's Kappa using the Statistical Package for the Social Sciences (SPSS), version 21.

## Results and Discussion

The result of the reliability analysis comparing human and machine-scoring was quite disappointing. The kappa statistic calculated between human raters was  $\kappa = .85$ . However, the inter-rater agreement between human and machine raters was calculated as  $\kappa = .16$  as shown in Table 6.1. This only demonstrated slight agreement between the human and machine raters. An analysis of the automated Concept Retrieval Technique was undertaken and four potential issues were identified that may help explain the low inter-rater agreement between the human and machine raters.

Table 6.1

*The values of Cohen's  $\kappa$  calculated for a secondary science class*

<i>Sample</i>	<i>n</i>	<i>Manual (<math>\kappa</math>)</i>	<i>Automated (<math>\kappa</math>)</i>	<i>Agreement</i>
Periodic Table	55	.85	.16	Slight

*Note:* The total number of students and kappa the reliability study conducted across for the topic “periodic table”.

Firstly, in the target word list for the “Periodic Table” (see Table 4.1) one of the targets for a correct mark was “any element example”. In this instance, a human rater with expert knowledge could make judgements upon the suitability of elements provided by test-takers. However, a machine can only score what has been inputted by the user and any decisions that are made are programmed by the developer. This was one reason why the inter-rater reliability was so low. It also provided feedback regarding the functionality of automated assessment and insights into the construction of appropriate target word lists. Furthermore, after a discussion with the subject-matter expert, it was revealed that most elements listed were not explicitly addressed in the learning episode and were generated from prior knowledge. The purpose of the assessment instrument is to assess the acquisition of conceptual knowledge from a specific learning episode. Therefore, the Concept Retrieval Technique for the topic “periodic table” was scored again by both human raters, adhering to the changes in the target word list. A new kappa statistic was calculated between human raters that was  $\kappa = .89$ . For future versions it is important to include a

prompt for users, that highlights to need to ensure the target concepts are specifically related to the concept trigger, and are addressed within a learning episode.

Secondly, analysing the machine-scored responses we found that some linguistic mistakes were made. Specifically, involving the use suffixes in the target concepts. For example, the concept *reaction* was scored as a target concept, which many students were able to successfully identify. However, some students missed out on a mark within the machine-scoring process if they used the term *reactivity* or *reactive*. The use of different suffixes is an issue that could be resolved by providing a shortened format of the word e.g., *react* instead of *reaction* or *atom* instead of *atomic*. However, this may not be feasible for all situations and will need to be addressed in version 2 to improve the overall reliability.

Thirdly, a number of spelling issues were identified in the test-taker responses that may have impacted on the inter-rater agreement. The human raters would be able to identify and make appropriate decisions on all incorrectly spelt concepts. For instance, words such as *nuetron*, *elemnt* and *proten* did not receive a mark in the machine-scoring. At present there is no spell-checking process available prior to machine-scoring with the automated Concept Retrieval Technique (Version 1). Therefore, it is important that this feature is added in the next version to improve the overall reliability. Also, the addition of the word cloud visualisation will provide the opportunity for users to identify commonly misspelt words and make adjustments with the target word list.

Finally, the scoring of duplicate responses was an issue that was identified in the analysis of the machine-scored responses. Especially, when students are forced to provide a response in all text boxes provided, they may duplicate a concept in their responses in order to submit the form. If the response that they duplicate is a target concept, there is the potential that could be awarded a mark multiple times for the single concept. Also, if the concept is embedded in other concepts they may receive multiple marks for the one concept. For example, *atom* is a target concept, however many test-takers also provided *atomic number* or *atomic mass*. In this instance a test-taker could receive three marks for the concept *atom*. This presents a major validity issue with the automation process, as human raters would identify this issue and only award one mark to the test-taker.

## 6.8 Summary of key findings

The achievement of design objectives is critical to the ongoing development of the automated Concept Retrieval Technique (Version 1). Firstly, the ability for test-taker data to be collected and stored was achieved as a third-party application was used to administer this task.



However, some issues were identified regarding the sustainability of this process as the data needed to be cleaned prior to importing into the online database to ensure fatal errors did not occur. Secondly, the ability for the user to create a dynamic online database was achieved. Users have complete autonomy regarding the size of the database, specifically the number of concepts that will be imported into the database. Some issues have been identified regarding adding additional fields to the database schema (e.g., Users are currently only allowed to have one identifier field, such as username). Therefore, if a user wanted to include first name and surname this could impact on the validity of the machine-scoring or have the potential to cause a fatal error. Considerations will be made for subsequent versions to address this issue. Thirdly, the ability for users to upload and store csv data files was achieved. However, some issues occurred in the administration of this process. For instance, users often inputted incorrect variables which caused significant run-time errors. For example, a Concept Retrieval Technique was administered and students were asked to provide 20 concepts related to a trigger. However, the user entered 15 in creating the online database for machine-scoring. This resulted in errors during the csv import as the data did not match the field columns. The next version will provide feedback to the user regarding this input and when the error occurs.

Finally, the operation of the scoring engine (e.g., search and scoring submodule) was achieved, but there are concerns regarding the overall reliability of these submodules. The results from study 6 suggest that some adjustments are needed in the next version to ensure that it is a reliable assessment instrument. Three issues need to be addressed in the automated Concept Retrieval Technique (Version 2) to improve the inter-rater reliability. Firstly, the issue pertaining to incorrect spelling needs to be addressed, especially if this assessment measure is to be applied to a primary and secondary education context. At present all incorrectly spelt concepts are not awarded a mark with the machine-scoring process. The scoring engine can only provide test-takers a mark when a response correctly matches a target concept, including correct spelling. This presents an issue with inter-rater reliability as a human rater would be able to identify and make appropriate decisions concerning spelling mistakes. It is recommended that inclusion of an automatic spell-checking submodule be implemented in the next version, clean up test-taker responses before the scoring engine is activated. This submodule will identify all possible incorrectly spelt words for the user, however they will not automatically address the issue. The user must select if a spelling error should be changed. It is anticipated that this will alleviate any issues with changing the integrity of the answer.

Secondly, the linguistic issues identified in study 6 with the use of a suffixes may prove difficult to resolve. The use of a prompt for the user when generating the target word list may minimize this issue. However, generating a visualisation of all terms entered by students in the form of a word cloud allows the user to make judgements regarding the targeted concept list, specifically issues with the use of suffixes. Both solutions will only attempt to minimize the issue. Further research is needed to investigate ways of utilising synonym submodules and providing that information to users in determining the target word list.

Finally, the issue with duplicate responses needs to be addressed. For example, Figure 6.24 provides a sample of the scored test-taker responses for the topic “periodic table”. The target concepts *atom* and *period* have been scored on multiple occasions. Specifically, in this example the test-taker will receive a total of six marks for two concepts as the program logic does not break and move onto the next student once the correct target concept is found. This issue would be a contributing factor in reducing the overall inter-rater reliability of the automated Concept Retrieval Technique (Version 1). Therefore, the changes in the next version will require enhancements to the search and scoring submodule to include a code break when a target concept is found in test-takers responses. It may be necessary to use a temporary data structure such as an array to contain correctly scored concepts or when concepts are scored they are deleted from the target word list, with the list being repopulated when the program moves to a new student. Furthermore, the programming logic search structure will need to be improved. Currently, focusing on test-taker responses one by one and comparing them to the entire target word list is not an efficient or effective process. The next version will address this issue.

Test Taker	Atom	Proton	Atomic Mass	Atomic Number	Period	Periodic Table	Form	Shell	Atomic Structures	Periods
------------	------	--------	-------------	---------------	--------	----------------	------	-------	-------------------	---------

Figure 6.24. An example of a student response to the solar cells Concept Retrieval Technique showing duplicate responses.



A background of a complex network graph with white and grey nodes connected by thin lines, overlaid on a light grey geometric pattern.

# 7

**Chapter**

# **Automated Concept Retrieval Technique (Version 2)**

## 7.1 Introduction

In the previous chapter, the construction of an automated Concept Retrieval Technique that utilised machine-scoring was achieved. However, the establishment of reliability evidence for the scoring engine revealed some inconsistency with the stability of the instrument, especially when compared to human raters. Slight inter-rater agreement was observed in the administration of the previous version, attesting to the potential of the automated Concept Retrieval Technique. However, a number of issues were identified and enhancements proposed to improve this result. Therefore, this chapter will present the next development of the automated Concept Retrieval Technique (Version 2), including an improved scoring engine that minimises the errors that occurred in the previous version. In addition, this version will include an opportunity for users to clean the inputted data to address errors, such as spelling mistakes. Furthermore, visualisations of test-taker responses will be generated to assist users, identify any consistent misconceptions or errors that have occurred in the administration of the test. In particular, concept visualisations will provide users with a snapshot of the test-taker learning and will inform any changes to the target word list required prior to automated scoring. These processes replicate the decisions that expert human raters make when scoring test-taker responses. Finally, the automated Concept Retrieval Technique (Version 2) will be subjected to further testing to determine if there is an improved inter-rater agreement between human raters and machine-scoring across a range of different test subjects and conditions.

## 7.2 Design objectives

The primary aim of the automated Concept Retrieval Technique (Version 2) is to improve the reliability of the automated scoring engine. In particular, the programming logic for the searching and scoring submodule will be reviewed to isolate the logic errors and improvements designed to alleviate the issue of scoring duplicate responses and improve the overall inter-rater reliability of the automated Concept Retrieval Technique (Version 2). Additionally, the ability to generate a visualisation of test-taker responses when the database is populated is an important improvement, allowing the user to examine the responses quickly and make informed decisions regarding scoring. This will assist users to detect commonly occurring spelling errors or the issues with the use of suffixes in test-taker responses. It is important that users apply this information in construction of the final target word list, prior to utilising machine-scoring.

Similarly, the ability to allow users autonomy in activating spelling changes. In the manual scoring of the Concept Retrieval Technique, decisions are made by the raters concerning the

spelling of test-taker responses. Concerns have been raised regarding the creation of such a submodule. The first concern is that an automated spell-checking submodule may change words significantly and alter their intended meaning. The second concern questions whether incorrectly spelling a concept, ultimately demonstrates conceptual knowledge. In view of these concerns, the spell-checking submodule must identify concepts that may be incorrectly spelt. However, any identified spelling error must not be changed unless the user accepts the change. This process will ensure that users have complete autonomy and discretion in marking incorrectly spelt concepts, simulating the scoring processes employed by a human rater. Given that, the following objectives have been identified to ensure the automated Concept Retrieval Technique (Version 2) improves the inter-rater reliability when compared to human raters.

### **7.2.1 Objective 1 – Improved search and scoring functionality of the scoring engine**

The first objective of the automated Concept Retrieval Technique (Version 2) is to improve the functionality of searching for target concepts and scoring these concepts, addressing the issue with duplicate concepts in test-taker responses. One solution identified in the previous chapter comprised of creating a temporary one-dimensional array to hold target word list concepts when they have been correctly scored. For each iteration of the search and score submodule the program would check if the target concept had been added to the temporary array. Next, if the target concept has not been scored previously, then it will be added to the temporary array and the test-takers score incremented. Finally, the temporary array will then reset every time the program progresses to the next test-taker. However, this solution requires excessive code and the use of data structures that could spawn new errors.

Another possibility involves reversing the programming logic for the search and scoring submodule. In the previous version, the programming logic was designed to check every concept in the target word list against every response submitted by a test-taker. Despite the efficient processing capabilities in managing excessive iterations, when the program moved to the next test-taker response it would analyse the entire target word list again. Given that, the proposed change in reversing the programming logic would start with the first concept in the target word list and compare it against all test-taker responses for the first test-taker. If a match occurs, then that test-takers score will be incremented, triggering the search process to move to the next test-taker in the database. As a result, this change will alleviate the scoring of multiple occurrences and reduce the need to use excessive code or data structures. In addition, if the target concept is not identified in test-taker responses then the search process will also move to the next test-taker and the process is

continued until all test-takers in the database have been exhausted. Finally, the process is repeated moving to the next concept in the target word list and the process is continued until all concepts in the target word list have been exhausted. Figure 7.1 provides an example of the first iteration of the enhanced searching and scoring submodule.

Target Concept List								
Atom	Proton	Mass	..	..	..	..	..	

↓

Test Taker 1	Atom	Formula	Period	Proton	Mass	Table	..	..	1
Test Taker 2	Shell	Atom	Mass	Atom	Element	Reaction	..	..	0
Test Taker 3	Shells	Promton	Shells	Atom	Atomic	Form	..	..	0
Test Taker 4	Protons	Protos	Mass	Periodic	Atomic	Period	..	..	0

Figure 7.1. The first iteration of the search and scoring submodule, where the first target concept (Atom) is checked against all first concept listed for test-taker 1.

The first iteration of the search and scoring submodule focuses on the first concept in the target word list (i.e., Atom). The search then starts with the initial response for test-taker 1 and a comparison is made between the two items (i.e., Does “Atom” match “Atom”). In the example above, this comparison is correct and the test-taker score is incremented. Because a correct concept was found the search shifts to the next test-taker (i.e., Test-Taker 2). Importantly, the focus is still on the first concept in the target word list, improving the efficiency of the programming logic within the search and scoring submodule. Figure 7.2 provides an example of the second iteration of the searching and scoring submodule logic.

Target Concept List									
	Atom	Proton	Mass	..	..	..	..	..	
	↓	↘	↘	↘	↘	↘			
Test Taker 1	Atom	Formula	Period	Proton	Mass	Table	..	..	Score 1 1 1 1
Test Taker 2	Shell	Atom	Mass	Atom	Element	Reaction	..	..	
Test Taker 3	Shells	Promton	Shells	Atom	Atomic	Form	..	..	
Test Taker 4	Protons	Protos	Mass	Periodic	Atomic	Period	..	..	

Figure 7.2. The second iteration of the searching and scoring submodule, where the search continues for each test-taker, focusing on the first target concept (Atom).

The searching and scoring submodule will continue iterating through the responses, effecting comparisons to the first concept in the target word list, until all test-takers in the database have been exhausted. When a comparison match is found, scores will be updated accordingly and the search focus will move to the next test-taker. For example, test-taker 2 has included the concept *Atom* on two occasions and in the previous version of the automated Concept Retrieval Technique they would have had that single concept scored twice, producing a significant effect on the inter-rater reliability of the scoring engine. However, breaking the search process when a correct concept is found, resolves this issue in the current version. Figure 7.3 provides an example of the third iteration of the searching and scoring submodule logic, where the process is repeated for the second concept in the target word list (i.e., Proton).

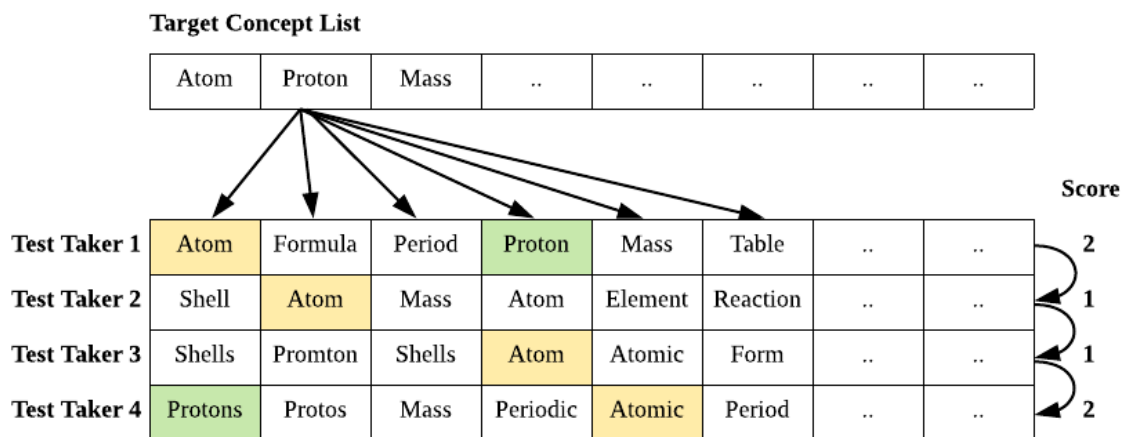


Figure 7.3. The third iteration of the searching and scoring submodule, where the process is repeated.

In the third iteration the focus will shift to the next concept in the target word list (i.e., Proton). At this point the same searching and scoring process will continue, starting back at the first response of test-taker 1. Again, the iterations will continue with the focus on the second concept in the target word list (i.e., Proton) until all test-takers in the database are exhausted. This third iteration highlights some improved capabilities in the functionality of the search and scoring submodule. For instance, the target concept (i.e., Proton) is not found in the responses of test-taker 2 and the search process moves to the next test-taker. Also, in searching the responses of test-taker 3, there is the occurrence of a concept spelling mistake (i.e., Promton). As a result, the concept is not scored, however further improvements are intended for this version that will provide the user with the opportunity to clean spelling mistakes upon the uploading of test-taker responses. Finally, both processes will continue until there are no more concepts in the target word list and the searching of all test-takers have been exhausted. Figure 7.4 provides an example of the final



iteration of the searching and scoring submodule with the final scores calculated for each test-taker.

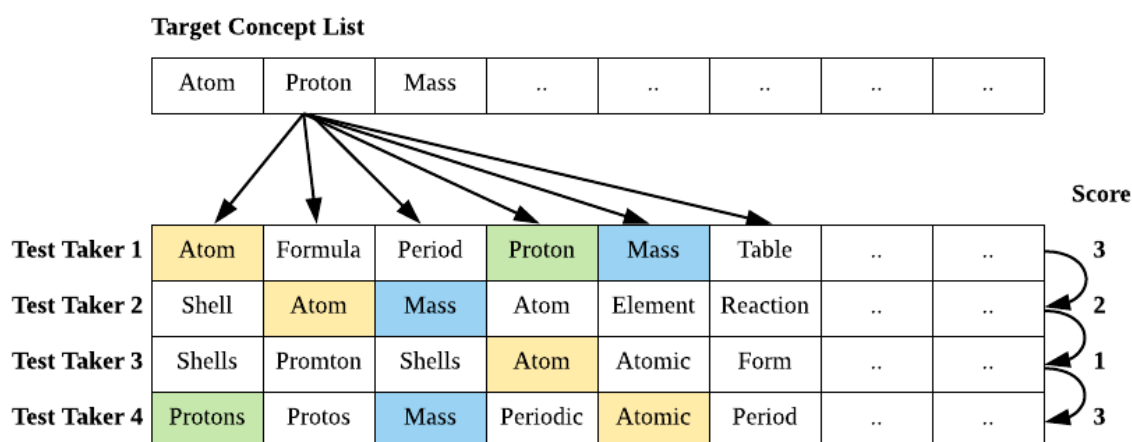


Figure 7.4. The final iteration of the search and scoring submodule, all target concepts have been scored for all test-takers.

### 7.2.2 Objective 2 – Generating word cloud visualisations of test-taker responses

The ability to create a word cloud visualisation of test-taker responses is an important feature of the new version, specifically for users to retrieve global information about the conceptual knowledge of the entire class. A word cloud is an easy to comprehend and quick to embrace visualisation, that has previously been used to analyse the use of vocabulary and grammatical tenses by uploading students writing into word clouds (Jayashankar & Sridaran, 2017). Firstly, the word cloud will highlight the common misunderstandings or errors (i.e., spelling) that may impact the reliability of the machine-scoring process. For instance, the construction of a word cloud for the Concept Retrieval Technique on the topic “periodic table” highlighted spelling issues (i.e., nuetrons) and suffix issues to be considered (i.e., atom, atoms and atomic). Secondly, the information provided in the word cloud may also substantiate and further inform the concept target word list. It would be envisaged that the majority of target concepts would be surfaced in the word cloud visualisation. Additionally, the word cloud visualisation will also serve as an evaluation tool of the learning episode. It would be expected that the concept objectives identified prior to the administration of the learning episode would be predominately present in the word cloud visualisation. Finally, this submodule will be activated when the user uploads a dataset to the online database.

### 7.2.3 Objective 3 – Data cleaning of test-taker responses

Misspelled words are common in student writing and have been identified as one of the most frequent errors in student free text responses. For example, Connors and Lunsford (1988) reported that approximately 25% of all errors found in a sample of 200 students' essays were misspelled words. Therefore, the ability for users to globally clean the test-taker responses to eliminate common errors is a necessary objective that is expected to have a positive effect on inter-rater reliability of the machine-scoring processes. However, some would argue that conceptual knowledge encompasses both the ability to retrieve a concept based on trigger and the capability to spell it correctly. Even so, the administration of the test in an online environment may encourage a greater frequency of spelling errors. The data cleaning module will simulate the processes employed by the rater in the manual marking process by identifying and rectifying simple spelling mistakes or inaccuracies. Secondly, it is important that the spell-checking submodule only highlights possible spelling errors and does not automatically alter the spelling. Flor and Fugati (2012) found in a study that automatic systems have significant limitations especially in the use of domain-specific words. Their results found that the automated program did not correct more than 20% of the spelling errors in the text and more than 20% of the correction suggestions by the program were incorrect. In addition, this could have a negative impact on inter-rater reliability if this process is not governed by the user. Therefore, all spelling errors will only be identified to the user and they will be required to select the errors to be fixed. This process will then change all occurrences of that spelling error within all test-taker responses. Finally, there are a number of third party spell-checking submodules that will be utilised to ensure objective 3 is achieved.

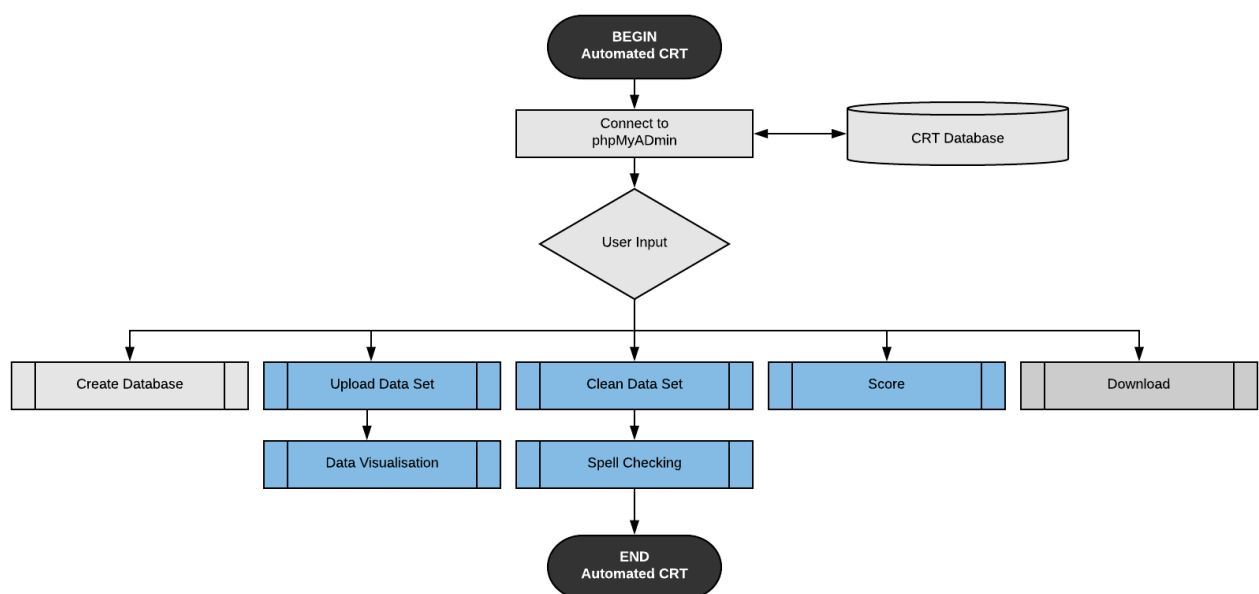
## 7.3 System modelling

The same system modelling tools that were utilized in version 1 will be employed to represent the automated Concept Retrieval Technique (Version 2). This version will employ some major and minor changes that will be modelled via system flowcharts. The system flowcharts will help gain a deeper understanding to how each version objective is achieved in the implemented software solution.

### 7.3.1 The automated Concept Retrieval Technique (Main Module)

The main module was constructed in the previous version, however significant improvements have been proposed in the current version, specifically the uploading, cleaning and scoring submodules. Figure 7.5 represents the system flowchart for the main module with all

submodules related to the current version isolated in the colour blue. Firstly, the upload dataset submodule was functional in the previous version, however this module will now be used to call on the data visualisation submodule. It is assumed that when test-taker data is imported into the online database the same data could be used to generate the data visualisation. This word cloud visualisation will determine the commonly occurring concepts from test-taker responses and display the result as an image within the user interface. Next, the clean dataset submodule will be constructed to identify and display possible spelling issues to the user. This submodule will call upon the spell-checking submodule to analyse test-taker responses and flag possible spelling errors, storing them in a temporary array. The array will then be displayed via the user interface to the user. Each identified spelling error will be listed with a proposed solution, requiring a decision from the user regarding which errors to rectify. In view that, this process simulates the actions that subject-matter expert would undertake in manual scoring of the Concept Retrieval Technique. Finally, the scoring engine will be significantly modified to improve the inter-rater reliability of the machine-scoring processes, specifically to address the issue with scoring duplicate responses. The changes applied to the programming logic of the scoring engine will ensure that it is more efficient and effective in searching for and scoring concepts from the target word list.



*Figure 7.5.* System flowchart of the automated Concept Retrieval Technique (Version 2) with the key submodules highlighted.

### 7.3.2 Upload dataset (Submodule)

The purpose of the changes identified for the upload dataset submodule are to address design objective 2. Figure 7.6 represents the system flowchart for the modified upload dataset

submodule, including the data visualisation submodule. This submodule is a third-party resource accessed within the MAMP server applications software infrastructure that will construct and output a word cloud. During the execution of the uploading dataset submodule it is called and a link made with the uploaded test-taker responses. This submodule will then count the commonly occurring concepts from test-taker responses. Consequently, this information will then be sorted in descending order and the first fifteen concepts will be displayed in the user interface via an outputted word cloud image.

To clarify, the word cloud image will present each concept in different colours to differentiate and the font size of concepts displayed will vary according to the frequency of the found concepts within the test-taker responses. For instance, in the Concept Retrieval Technique for the periodic table it would be expected that commonly occurring concepts (e.g., Atom and Mass) would occur frequently in test-taker responses. As a result, these concepts would appear in the largest font within the word cloud visualisation. The only constraint in this submodule is that the maximum number of concepts to be displayed in the word cloud is restricted to a static input. Currently, it is a maximum of fifteen concepts, which ensures that the visualisation is coherent and not confusing by displaying an extensive number of concepts. On the other hand, for Concept Retrieval Techniques that require a large retrieval of concepts it may not be appropriate to limit the visualisation to fifteen concepts. Therefore, in future versions the user should be able to determine the maximum number of concepts to be displayed within the visualisation.

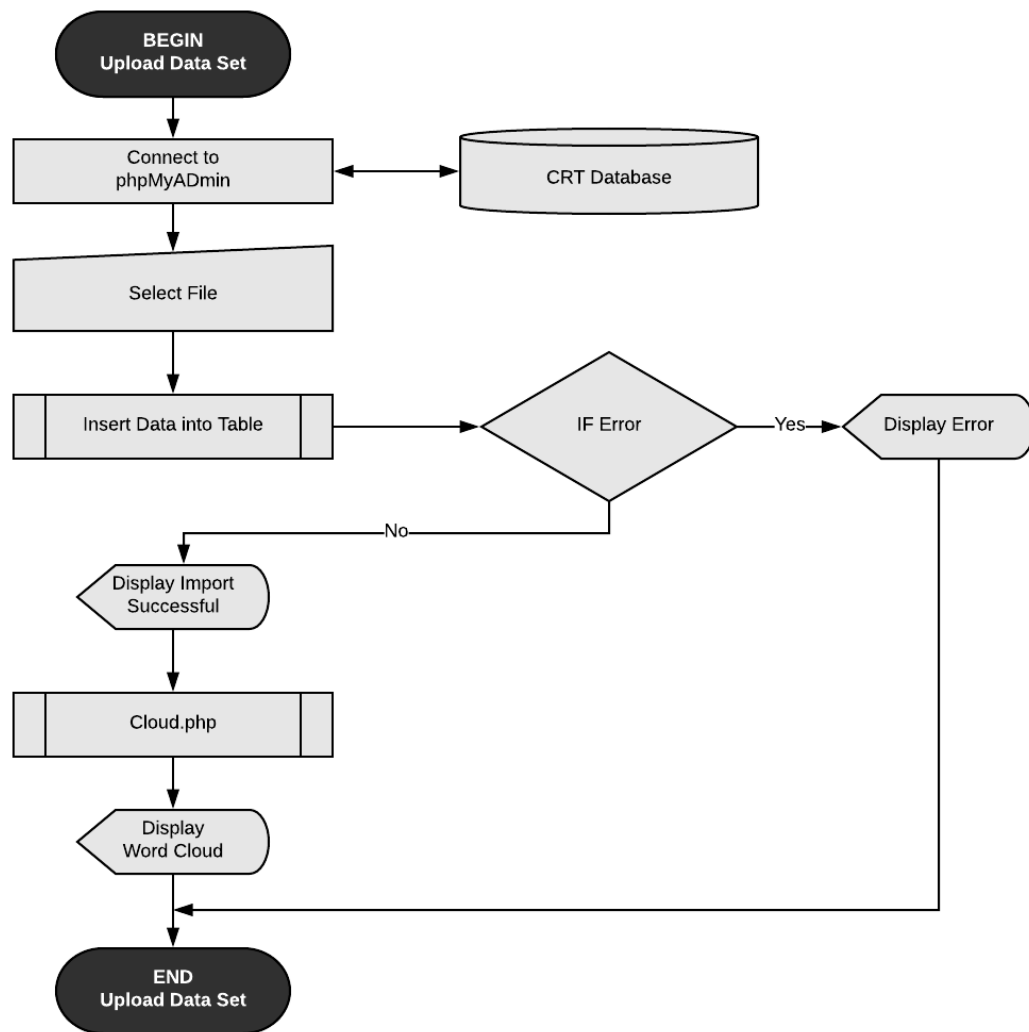


Figure 7.6. System flowchart of the revised upload dataset submodule to include word cloud data visualisations.

### 7.3.3 Clean dataset (Submodule)

The purpose of the clean dataset submodule is to address design objective 3. This submodule was not constructed in the previous version. Figure 7.7 represents the system flowchart for the clean dataset submodule that calls on the spell-checking submodule once the test-taker responses have been uploaded into the online database. The spell-checking submodule is a third-party resource accessed within the MAMP server applications software infrastructure. This submodule will sequentially search test-taker responses for the occurrence of identified spelling errors. Every word will be analysed to determine spelling issues and if an issue is discovered a spelling suggestion will be determined. Both the error and suggestion will be stored in a temporary array that will be sent back to the clean dataset submodule for output. The user interface will then

export the temporary array of spelling errors and suggestions with a command button aligned to each error. The user will then be required to select the errors to be actioned by clicking on each command button aligned to the spelling error.

Finally, all selected spelling issues will be addressed within test-taker responses in the online database, simulating the processes applied by a manual rater in scoring the Concept Retrieval Technique, enabling an improved inter-rater reliability. However, there are a number of potential errors that can occur from the use of automated spell-checking submodules. Firstly, these systems cannot infer the true meaning of misspelt words or spacing errors with this misclassified data having the potential to hinder the machine-scoring process (Flor & Fugati, 2012; Muhlenbach et al., 2004). Secondly, there is the possibility that the language type (i.e., US English vs. UK English) may cause significant issues in the implementation of this submodule. Hence, users may need to select the type of language to be used by the spell-checking submodule to alleviate the frequency of these issues, which may be an enhancement in the next version.

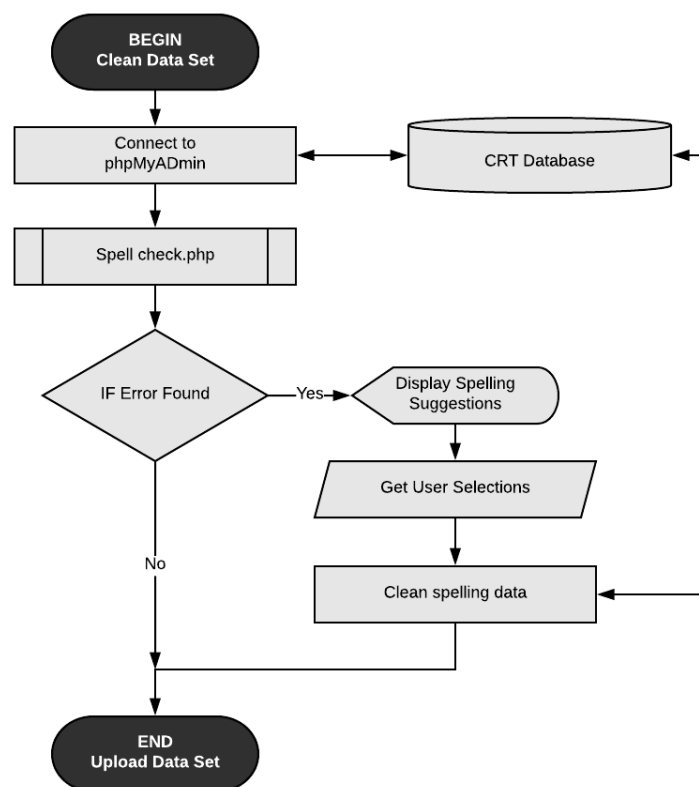


Figure 7.7. System flowchart of the clean dataset submodule to include the spell-checking submodule.

### 7.3.4 Scoring engine (Submodule)

The purpose for the modification of the scoring engine submodule is to address design objective 1. This design objective focuses on improving the overall inter-rater reliability of the automated Concept Retrieval Technique (Version 2) by eliminating the scoring of duplicate concepts within test-taker responses. However, there are no significant changes to the system flowchart represented in the previous version. The significant changes occur within the programming logic, not the flow of processes. Therefore, the changes to this submodule will be represented in the algorithms generated in the submodule construction.

## 7.4 Module construction

The algorithms below have been provided in this thesis to help understand the programming logic and how the objectives have been achieved in the creation of the automated Concept Retrieval Technique (Version 2). Each of the algorithm flowchart provided will assist in providing a deeper understanding of the different processes, decision and iterations that are used to achieve the current version objectives.

### 7.4.1 Upload dataset (Submodule)

The upload dataset submodule has been modified to include the word cloud visualisation of the most commonly used concepts from test-taker responses. This submodule constructed in the current version was already successful in accessing the selected csv data file and uploading the user responses for each test-taker into unique fields within the online database table. However, the aim of the automated Concept Retrieval Technique (Version 2) is to include the calling of the cloud.php submodule to count the frequency of commonly used concepts within test-taker responses. This is achieved by sorting the concepts in alphabetical order and the frequency of each concept is then counted. Finally, the concepts are sorted again in descending order, according to the frequency count and the fifteen most commonly used concepts are used to populate the word cloud visualisation. In the next version there will be the opportunity for the user to increase the maximum size of the concepts captured in the visualisation and also the opportunity to change the font and text colours of the visualisation. However, in this version the primary aim is to achieve the creation of a word cloud and these elements will remain static. Figure 7.8 provides a representation of changes to the upload dataset submodule for the automated Concept Retrieval Technique (Version 2).

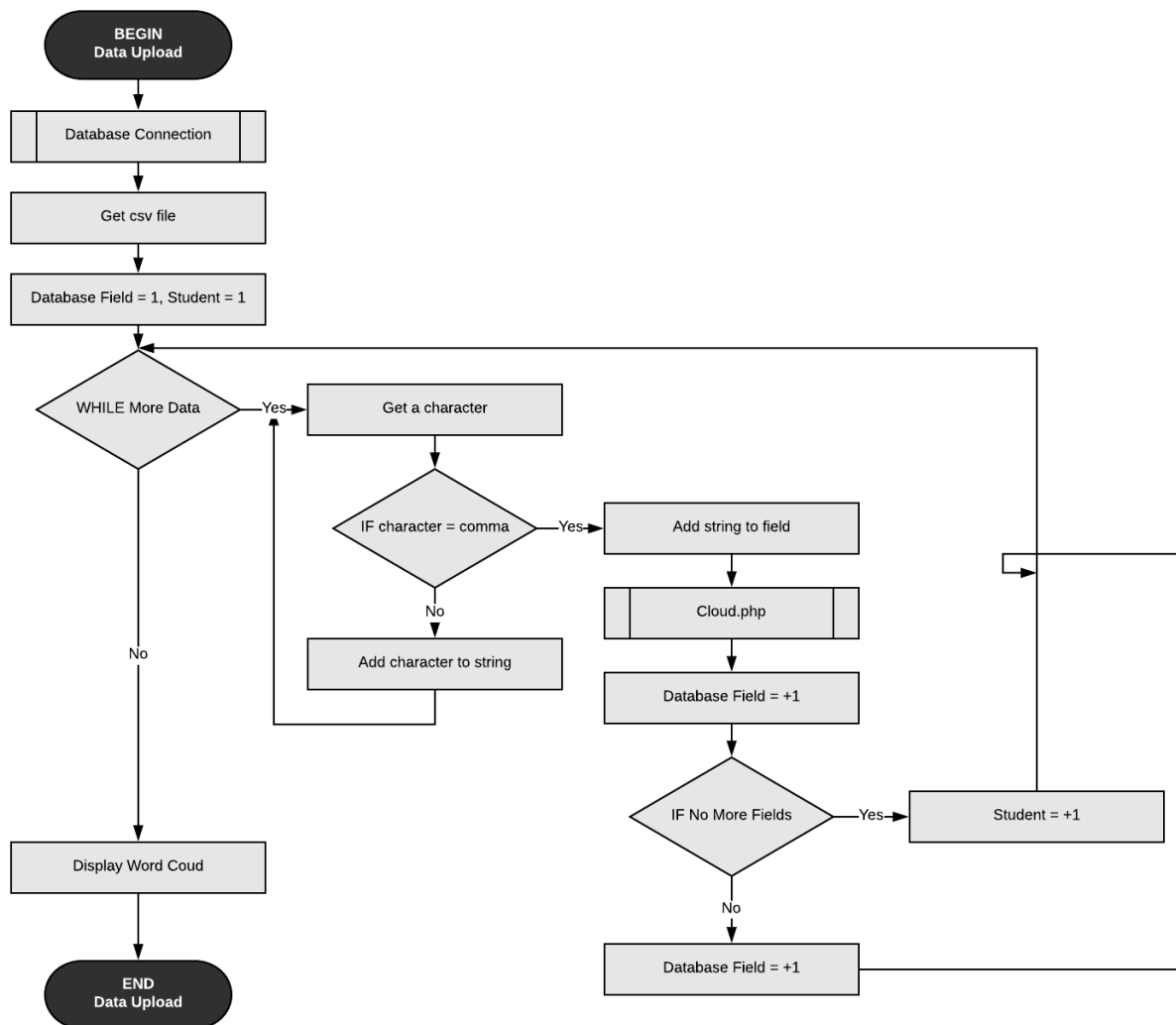


Figure 7.8. Algorithm representation for the upload dataset submodule with the word cloud visualisation.

The cloud.php submodule is a third-party resource that is called during the upload dataset submodule. Figure 7.9 provides the programming logic for constructing a word cloud visualisation. Firstly, all strings are stored separately in an array (e.g., Concept Array) until there are no more strings to be inputted. Secondly, the array of strings is sorted in alphabetical order. Thirdly, the array of strings is sequentially searched comparing each concept for a match and updating a counter aligned to each individual concept. If a match does not occur the counter is stored next to the concept. The aim is to count the frequency of a concept occurring extracted from test-taker responses. This process will continue until the list of concepts has been exhausted. Finally, the array is sorted again, according to the concept counter from highest to lowest. The first fifteen concepts within the array are used to generate the word cloud image that will be displayed to the user. The font size will depend on the differences in frequency between each concept. Table



7.2 provides an example of the information used to generate the word cloud. In this information, the difference in frequency between concept one and two is significant and is reflected in the font size (i.e., The font size for concept1 is 42 compared to concept2 which is 22). Finally, the colour selection is random, however this is something that could be modified in the next version. Figure 7.10 provides a representation of the interface for the upload dataset submodule with a word cloud visualisation generated from the Concept Retrieval Technique for the “periodic table”.

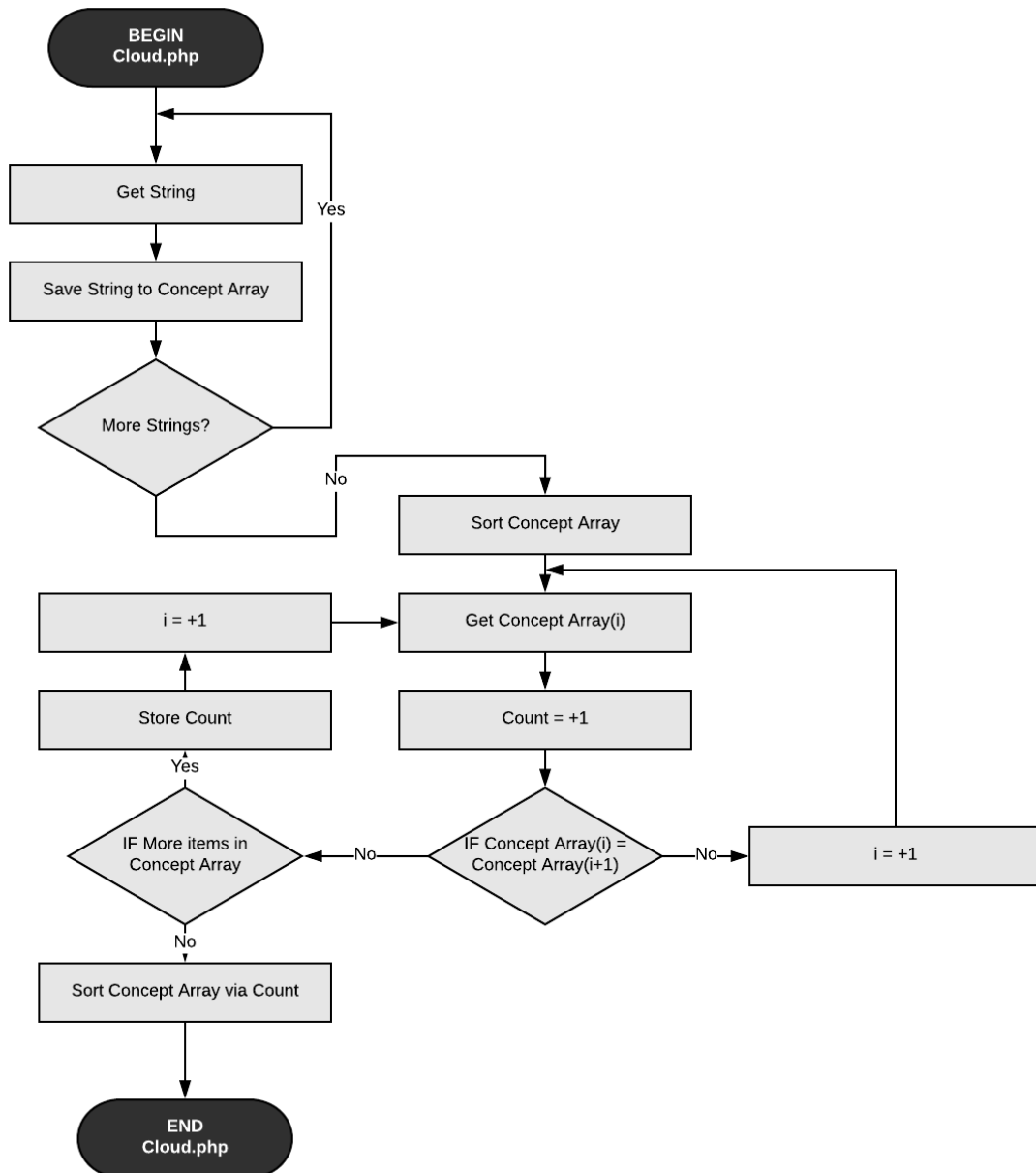


Figure 7.9 Algorithm representation for the cloud.php submodule.

Table 7.2

*An example of the information used to generate the word cloud visualisation.*

Concept Num	1	2	..	14	15
Count	32	17	..	4	3
Font Size	42	22	..	8	6
Colour	Green	Red	..	Blue	Yellow

**Concept Retrieval Test**
[Home](#)
[Create](#)
[Upload](#)
[Clean](#)
[Score](#)
[Download](#)
[Help](#)

## Upload Concepts

### Upload Student Data

Please select and upload your student data via csv file format.

No file chosen

### Concept Visualisation

You database has imported successfully.

*Figure 7.10.* The Interface for the upload dataset submodule with the word cloud visualisation with test-taker responses uploaded and visualisation created from the Concept Retrieval Technique for the “periodic table”.

#### 7.4.2 Clean dataset (Submodule)

It is predicted that the cleaning of the inputted data will only have a slight improvement on the inter-rater agreement of the automated Concept Retrieval Technique (Version 2) as the probability of spelling mistakes is quite low and commonly occurring mistakes will be identified from the visualisation word cloud. However, the clean dataset submodule will utilise the spell-

checking resources active within the MAMP server applications software infrastructure. The key consideration is that spelling errors will be identified for the user, they will not be automatically changed. This allows the user to judge the integrity of the change, simulating the process a manual rater would apply would during the scoring of the Concept Retrieval Technique. If the user selects to address a spelling issue, all instances of that issue in the database will be rectified. This process is repeated until the user has addressed all the spelling issues identified from the test-taker responses. In future versions, there will be the opportunity for users to change the default language, identify linguistic issues such as suffixes and allow the user to turn on or off the spell-checking submodule. Figure 7.11 provides an algorithm representation of the clean dataset submodule. Finally, the interface for the clean dataset submodule has been provided in Figure 7.12, which has diagnosed a number of spelling issues from the Concept Retrieval Technique on the “periodic table”.

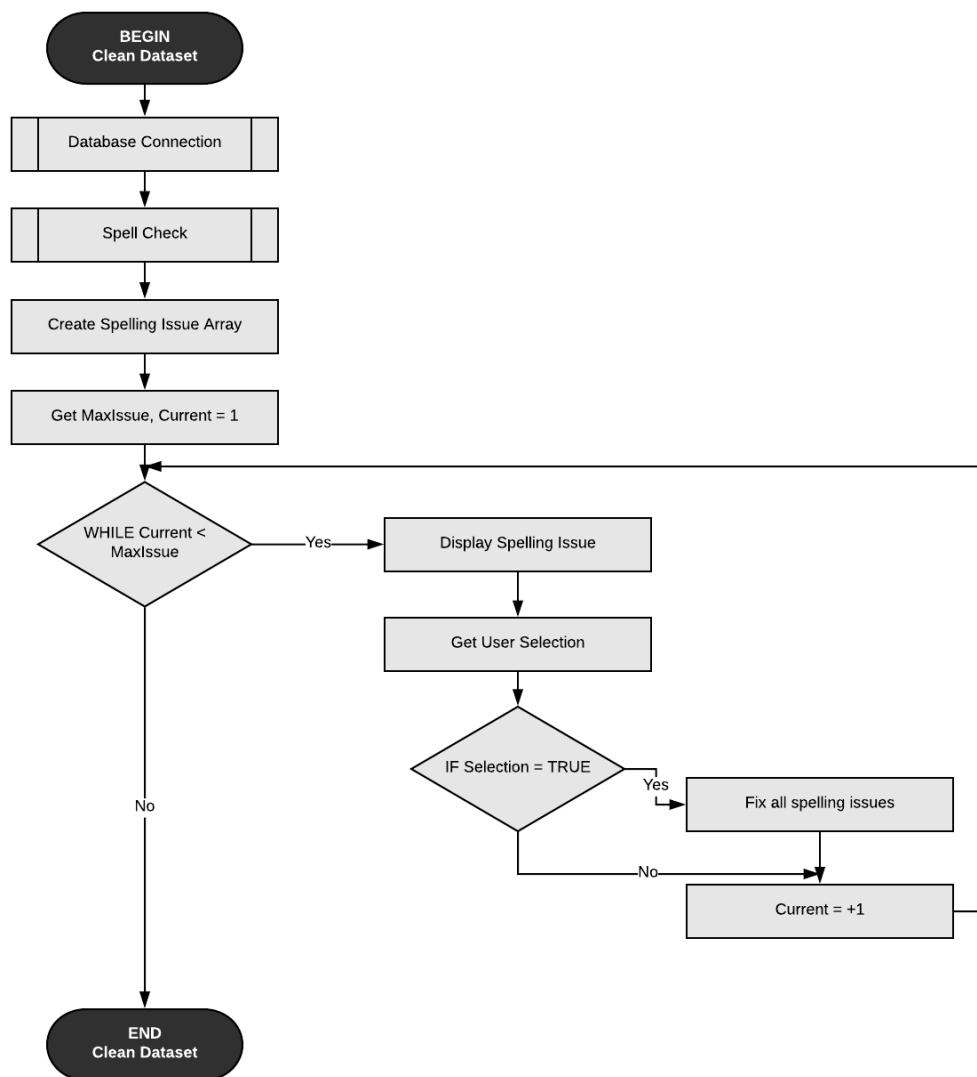


Figure 7.11. Algorithm representation clean dataset submodule.

**Concept Retrieval Test**
Home
Create
Upload
Clean
Score
Download
Help

## Clean Data

**Perform Spell Check**

Check Spelling

**Spelling Issues Identified**

Please select the spelling issues you would like to fix

Alkalimetal: ☐ Fix Spelling
Neutrons: ☐ Fix Spelling
Magnesium: ☐ Fix Spelling

Valency: ☐ Fix Spelling
Reactivity: ☐ Fix Spelling
Alkaline: ☐ Fix Spelling

Neutritisation: ☐ Fix Spelling
Alkaline: ☐ Fix Spelling
Neutrons: ☐ Fix Spelling

Properties: ☐ Fix Spelling
Electons: ☐ Fix Spelling
Valencies: ☐ Fix Spelling

Figure 7.12 Interface for the clean dataset submodule with test-taker responses uploaded from the Concept Retrieval Technique for the “periodic table”.

### 7.4.3 Scoring engine (Submodule)

The issues with duplicate responses have been addressed with the changes in programming logic applied to the scoring engine submodule. The user now has the ability to select the size of the target word list rather than being limited to ten possible target concepts. Figure 7.13 provides the algorithm representation for this submodule. The key difference with this algorithm is that the logic is built around iterating through the concept target word list and making comparisons with test-taker responses. Starting with the first concept in the target word list (e.g., TargetNum = 1) the algorithm then starts with first test-taker (e.g., Student = 1). Next, the first response is initialised (e.g., ConceptNum = 1) and the comparison statements are executed. If there is no match the algorithm increments to the next response submitted by the test-taker (e.g., ConceptNum = +1). This will iterate until the loop is broken by a concept match or the ConceptNum value exceeds the maximum number of responses submitted by the test-taker. When a match is found the score will increase (e.g., Student.Score = +1) and ConceptNum will be set to the maximum number of concepts (e.g., ConceptNum = MaxConcept). This statement breaks the loop and ensures duplicate concepts are not scored for each test-taker.

The process will continue for the next test-taker (e.g.,  $\text{Student} = \text{Student} + 1$ ), until all students in the database have been exhausted (e.g., The student variable value is greater than  $\text{MaxStudent}$ ). The algorithm will then focus on the next target concept (e.g.,  $\text{TargetNum} = +1$ ) and the process is repeated. Each time comparisons are made for all test-taker responses. The algorithm repetition will conclude once the  $\text{TargetNum}$  value exceeds the  $\text{MaxTarget}$  value, demonstrating that there are no more targets to be scored. Within the interface there is the opportunity for users to enter the maximum number of targets and the target word list is submitted as single line of text with each target separated by a comma. Figure 7.14 provides an example of the “periodic table” being scored with the interface for the scoring submodule, showing the changes to the user interface design.

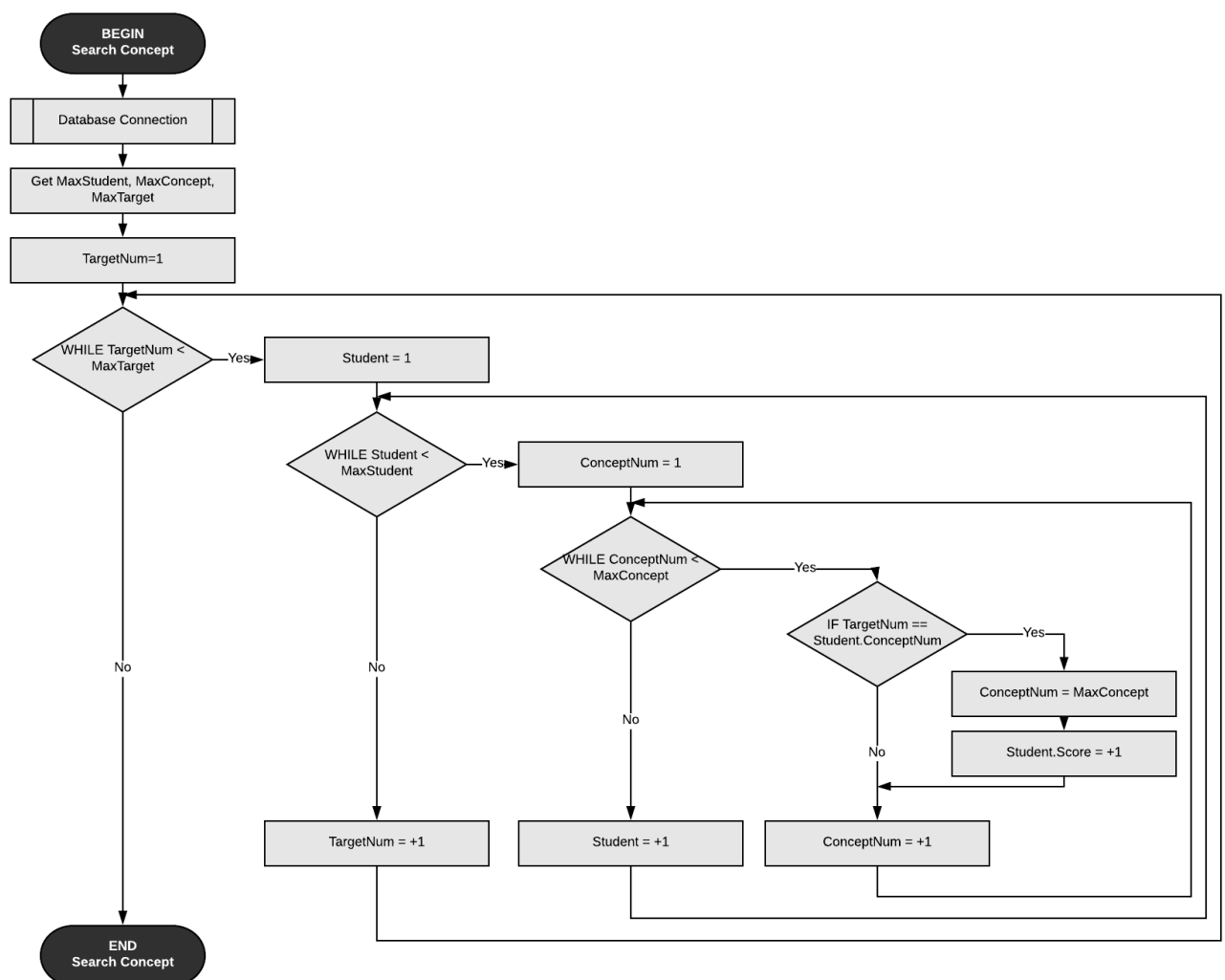


Figure 7.13. Algorithm representation for the updated search and score submodule.

**Concept Retrieval Test**[Home](#)[Create](#)[Upload](#)[Clean](#)[Score](#)[Download](#)[Help](#)

## Scoring Engine

**STEP 1: Number of concepts in the target list**

3

**STEP 2: Concept target list**

Please ensure every concept is seperated by a **comma**

Atom, Proton, Mass

Score CRT

Figure 7.14. Interface for the search and score submodule with an example data et from the topic “periodic table”.

### 7.5 Study 7 - Reliability of the automated Concept Retrieval Technique (Version 2)

In this study we used the same data as in study 5 to explore the improvements in the inter-rater reliability of the automated Concept Retrieval Technique (Version 2). The primary purpose of this study was to determine the improvement in overall reliability due to advancements implemented in the current version. We used Cohen’s kappa as a statistical measure to examine the correlation that indicates agreement and correspondence between machine-scores to human scores. Furthermore, this measure provided a determination of the inter-rater reliability and overall utility of the automated assessment measure. The student concepts from a previously scored Concept Retrieval Technique were imported into the automated Concept Retrieval Technique (Version 2) and a total score was generated for each student. Finally, the results of two independent raters were used to generate a total score and the results were then used to determine the consistency of the inter-rater agreement between the machine-scores and human scores.

## Method

### Participants

Fifty-four secondary school science students participated in the study from an all-male secondary school in Australia. Their average age was 15 years ( $SD = .52$ ). In test two, forty-five secondary school science students participated from the same school. Their average age was 13 years ( $SD = .32$ ). Participation was voluntary and students were not compensated for their participation.

### Materials

A Concept Retrieval Technique was developed for the topic of the “periodic table”. A subject-matter expert worked together with the researchers to generate the target word list containing all admissible concepts for the topic. All aspects of this study are detailed previously in study 3 & 5 of this thesis.

### Procedure

The administration of the Concept Retrieval Technique included an instruction, requiring students to write down all concepts or ideas about a topic using only keywords or bullet points, with the directive to avoid writing in full sentences. Following the administration of the Concept Retrieval Technique, two raters independently marked the responses with the help of the target word list. One mark was awarded for each correctly identified concept and a final score for each student was thereby generated. The inter-rater agreement generated from the two raters for the Concept Retrieval Technique was  $\kappa = .89$ . The raters then resolved any inconsistencies in their scores to generate a final score. This score was used as a measure of inter-rater reliability against the scores generated from the automated Concept Retrieval Technique (Version 2). The operation of the automated Concept Retrieval Technique (Version 2) required the same list of student concepts to be imported into the online database. Furthermore, the same target word list used by human raters was imported machine-scoring executed. Finally, the total scores of the machine-scoring process were downloaded and added to SPSS. Comparisons were made between both methods of assessment.

### Analysis

Inter-rater agreement was established by computing the Cohen’s Kappa using the Statistical Package for the Social Sciences (SPSS), version 21.

## Results and Discussion

The reliability analysis results for the machine-scoring of the Concept Retrieval Technique (Version 2) presented significant improvements from the comparisons made in the previous version. The reviewed kappa statistic calculated between human raters was equal to  $\kappa = .85$ . Given that, the inter-rater agreement calculated between human and machine raters for the automated Concept Retrieval Technique (Version 2) was calculated at  $\kappa = .70$  as shown in Table 7.3. This result demonstrated substantial inter-rater agreement and supported the increased stability of the automated scoring process. Furthermore, the high agreement in scores demonstrated by the kappa value, provided evidence that the machine-scoring process of the current version is indeed reliable as there is much less disagreement between human and machine-scoring processes than in version 1.

Table 7.3

*The values of Cohen's  $\kappa$  calculated for a secondary science class*

<i>Sample</i>	<i>n</i>	<i>Manual (<math>\kappa</math>)</i>	<i>Automated (<math>\kappa</math>)</i>	<i>Agreement</i>
Periodic Table	55	.85	.70	Substantial

*Note:* The total number of students and kappa for the reliability study conducted for the topic “Periodic Table”.

In summary, two conclusions can be drawn from the results of this study. Firstly, the high reliability coefficient found in the analysis of the automated Concept Retrieval Technique (Version 2) for the topic “periodic table” acknowledges the improvements implemented in the current version. Secondly, the results of this study highlight the need to assess the generalizability of the instrument across a broad spectrum of subjects as a reliable tool for measuring conceptual knowledge. A generalizability study establishes the degree of correlation of the instrument with other testing situations that assesses a similar construct. The evidence of the generalizability of the automated Concept Retrieval Technique (Version 2) is the opportunity to analyse the consistency of the human vs. machine-scoring processes across different subjects, students and testing conditions. Therefore, the next study will examine the reliability of the machine-scoring processes across different Concept Retrieval Technique implementations.



## 7.6 Study 8 – Generalizability of the Automated Concept Retrieval Technique (Version 2)

In Study 8 we explored the stability of the kappa when the automated Concept Retrieval Technique is administered across different school subjects and student age groups. The primary aim was to measure consistency in application and identify other potential issues to be addressed in the next version. For the automated Concept Retrieval Technique (Version 2) to be generalizable, it has to be tested across different conditions, which in this case would be the same conditions as the inter-rater agreement study. The results of Concept Retrieval Techniques administered in three different secondary science settings for topics motion, infectious diseases and solar cells were used and their inter-rater reliability measured. The consistency and overall stability in the kappa will be interpreted as evidence for the generalizability of the machine-scoring and support the improvements implemented from the previous version.

### Method

#### Participants

In this study, 45 secondary school science students participated (100% male,  $M$  age = 14 years,  $SD = .52$ ), 54 secondary school science students (100% male,  $M$  age = 14 years,  $SD = .52$ ) and 53 secondary school science students (100% male,  $M$  age = 14 years,  $SD = .52$ ). Participation was voluntary and the students did not receive any compensation for their participation.

#### Materials

A Concept Retrieval Technique was developed for all topics required subject-matter experts to work together with the researchers to generate target word list containing all admissible concepts for this topic. The target word lists used in all three topics are shown in Tables 4.1, 7.4 and 7.5 respectively.

Table 7.4

*The target word list used for scoring the topic motion*

<u>Admissible concepts</u>	
Gravity	Mass
Variable	Inertia
Acceleration	Newton
Force	Velocity

Momentum	$F=ma$
Friction	Speed

Table 7.5

*The target word list used for scoring the topic solar cells*

<u>Admissible concepts</u>	
Cell	Power
Voltage	Temperature
Series	Energy
Parallel	Photovoltaic

### Procedure

The administration of the Concept Retrieval Technique included an instruction, requiring students to write down all concepts or ideas about a topic using only keywords or bullet points, with the directive to avoid writing in full sentences. After the administration of the Concept Retrieval Technique, two raters independently marked the responses with the help of the target word list. One mark was awarded for each correctly identified concept and a final score for each student was thereby generated. The final scores generated by each rater was averaged and used as measure of inter-rater reliability against the overall scores of the automated Concept Retrieval Technique (Version 2). The same list of student concepts were imported into the automated Concept Retrieval Technique database. The same target word lists used by human raters were imported into the automated system and machine scored. Finally, the total scores of the machine-scoring process were added to SPSS as column to be compared against the scores of the human raters.

### Analysis

Inter-rater agreement was established by computing the Cohen's Kappa using the Statistical Package for the Social Sciences (SPSS), version 21.

### Results and Discussion

To determine the overall reliability of the instrument, three datasets were subjected to reliability analysis between human and machine raters. The results of this reliability analysis carried out for three Concept Retrieval Techniques across different subjects and age groups are

depicted in Table 7.6. Inspecting the kappa values, one can see that the inter-rater agreement has made significant improvements from the previous version and across different administrations it remains stable and reliable. The high agreement in scores suggest that the agreement in scores provided by the human and machine raters did not occur by chance. Furthermore, this evidence continues to support the automated Concept Retrieval Technique as reliable instrument that can be implemented across different contexts with minimal disagreement.

Table 7.6

*The values of Cohen's  $\kappa$  calculated across three different samples*

<i>Sample</i>	<i>n</i>	<i>Manual (<math>\kappa</math>)</i>	<i>Automated (<math>\kappa</math>)</i>	<i>Agreement</i>
Infectious Diseases	45	.91	.63	Substantial
Motion	54	.68	.68	Substantial
Solar Cells	53	.76	.81	Substantial

*Note:* The total number of students and kappa values for the reliability studies conducted across three different subjects and age groups.

It was observed among the three datasets, that the reliability between human and machine-scoring was still significant in terms of inter-rater agreement. However, the secondary school science topic of “infectious diseases” had the most significant difference in kappa value from .91 between the manual raters to  $\kappa = .63$  when compared to the machine-scoring. The cause of this difference can be found in the creation of the target word list. In this example, one of the target concept is any disease (e.g., AIDS or Ebola). Specifically, a human rater has the ability to make inferences regarding the scoring of only one concept that matches the criteria. However, machine-scoring must adhere to specific programmed criteria. Therefore, the automation can only score either one concept or both. In this study both concepts were entered into the target word list and resulted in the difference in inter-rater agreement.

The other two datasets provided consistent alignment between the human and machine raters and demonstrated minimal difference in inter-rater agreement. Given that, an analysis of the results was conducted and issues with spelling mistakes and linguistic issues such as the use of suffixes were still observed. For example, in the automated Concept Retrieval Technique (Version 2) for the secondary school science topic of “solar cells” there were issues with the spelling of *parallel* and the abbreviations for the term *photovoltaic*. In particular, the term *photovoltaic* was

abbreviated to *PV* by some students and this was scored as a correct answer by the human raters, but not in machine-scoring. Furthermore, the spelling mistakes were identified, however the changes were not implemented, so enhancements will need to be made in the next version with a spell-checking submodule.

## 7.7 Summary of key findings

The ongoing achievement of all design objectives is paramount to the development of the automated Concept Retrieval Technique (Version 2) and improvement of the inter-rater reliability. Firstly, the improved search and scoring functionality of the scoring engine resulted in an increase of inter-rater reliability from slight to substantial. Minimal issues were identified in this submodule that need to inform the developments of the next version. However, some errors were identified in the uploading of different test-taker responses in study 8. Specifically, in the motion and solar cells dataset, if the user responses contained redundant characters (i.e., comma or apostrophe) this would result in a runtime error. Figure 7.15 provides a screenshot of the error message and upon investigation it was found that the import dataset submodule had difficulty in handling these characters as a result of the characters representing other actions in the program. Therefore, modifications will need to be made in the next version to ensure that the upload dataset submodule is stable in handling redundant characters from user responses.

database error: You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near '","-State of matter',"-Melting point',"-Heat and electrical conductivity',"-' at line 1

*Figure 7.15.* The error message generated from the importing of a test-taker response dataset containing redundant characters.

Secondly, the ability to generate word cloud visualisations from test-taker responses was achieved in this version. These visualisations correlated with the initial target word lists. However, feedback from the operation of this version by users found that it was difficult to infer the differences between the concepts in the visualisation (i.e., different sizes and colours). This may be a result of the limited number of concepts selected in the visualisation or the colour scheme used by the interface. As a result, in version 3 users would benefit from the opportunity to set the maximum number of concepts in the word cloud visualisation. However, considerations need to be made regarding the size of the space allocated within the user interface to display the visualisation. Specifically, if a user selected a variable (i.e., 100) it may be difficult to produce

such an image in the space available in the interface. Further testing of this feature will be done to determine the suitability. Finally, the opportunity for the user to adjust the colours and fonts in next version will be explored in the next version. However, this is primarily for aesthetic purposes and will not have any impact on the inter-rater reliability of the automated Concept Retrieval Technique.

Thirdly, the data cleaning of the test-taker responses was mostly achieved by utilising the spell-checking submodule. The current version of this submodule only provides the user with a list of spelling errors and an opportunity to fix the spelling issue by clicking on a command button. If the user is not aware of the proposed change they may make a change that will have a significant effect on the integrity of a user's response. Therefore, in the next version the data cleaning submodule will need to address the issue of providing users with the spelling recommendations. Also, as stated previously it will be beneficial to allow the user the opportunity to select the language used by the spell-checking submodule.

A background graphic consisting of a complex network of interconnected nodes and lines. The nodes are represented by small circles in white, dark gray, and light gray. The lines are thin and connect the nodes in a web-like pattern, with some lines being white and others dark gray. The overall effect is a sense of a large, interconnected system or network.

# 8

**Chapter**

# **Automated Concept Retrieval Technique (Version 3)**

## 8.1 Introduction

In the previous chapter, the automated Concept Retrieval Technique (Version 2) was enhanced to improve the stability of this instrument, especially in comparison to human raters. Substantial inter-rater agreement was obtained, verifying the effectiveness of the automated Concept Retrieval Technique. The previous version successfully addressed a number of issues regarding the performance of the machine-scoring process. Even so, there are some more issues that need to be addressed to enhance the overall suitability of the software program and minimising the impact of user errors, especially when beta testing is undertaken. Therefore, this chapter will present the next development of the automated Concept Retrieval Technique (Version 3), including enhanced data cleaning from the import dataset submodule and providing more user capabilities to the data visualisation and spell-checking submodules. In addition, this version also includes the opportunity for users to download test-taker responses with the total score calculated. This information provides users with the important formative assessment insights that can inform subsequent learning episodes. Furthermore, ensuring that the downloaded dataset is in a recognisable format (i.e., csv) will make sure the information is agile and can be imported into other systems for further manipulation. Finally, the automated Concept Retrieval Technique (Version 3) will be moved from the local environment where alpha testing has been conducted by the programmer, to an online environment where beta testing can be conducted by users.

## 8.2 Design Objectives

The primary aim of the automated Concept Retrieval Technique (Version 3) is to improve the overall functionality of the program to minimise the occurrence of user generated errors, especially when the program is moved from a local to an online environment. Furthermore, the program must allow users to download the scored responses to inform future learning episodes. Therefore, the upload dataset and data cleaning submodules will be enhanced to demonstrate increased stability to withstand user beta testing. Furthermore, the download dataset submodule will be constructed, allowing users to download datasets in the csv format. Finally, all user instructions within the interface and other help resources will be reviewed to ensure that user errors are minimised and the instrument is stable for beta testing. Given that, the following objectives have been constructed to ensure the automated Concept Retrieval Technique (Version 3) improves the overall functionality in preparation beta testing in an online software environment.

### **8.2.1 Objective 1 – Downloading the scored test-taker responses**

The first objective of the automated Concept Retrieval Technique (Version 3) is to allow users the ability to download the scored test-taker responses. In this version, the only format for users to download will be the csv format. This format is malleable and predominately used in importing data between software systems. This objective will require a connection to the online database and the extraction of relevant data in the csv format. With the current format all fields will be provided in the dataset download including CRTID, test-taker responses and scores.

### **8.2.2 Objective 2 – Enhancing the upload dataset to remove redundant characters**

The previous version highlighted some issues with the presence of redundant characters in the test-taker responses and runtime error caused upon import. The stability of the program in online environment is a concern, especially for beta testing as these types of errors need to be identified and appropriately fixed. Presently, the comma and apostrophe are the only characters causing an error. However, other characters will be tested and feedback provided regarding how they are handled. Furthermore, if an error does occur there needs to be some form of user feedback provided to help the user to navigate the error, rather than the database error provided in the previous version.

### **8.2.3 Objective 3 – Allowing user interactivity in the construction of the data visualisation**

The ability to generate word cloud visualisations from test-taker responses was achieved in the previous version. However, feedback suggest that its overall effectiveness would be increased if users could increase the number of concepts beyond the current maximum of 15 and the ability to adjust the colours and fonts in the visualisation. To achieve this objective the automated Concept Retrieval Technique (Version 3) will have an input text box for users to enter the maximum number of concepts to be visualised. This will limit user input error and ensure that a variable is not selected that will jeopardise the screen space available for the visualisation. Finally, users will be able to select a different colour scheme randomisation, which will change the output of the visualisation.

### **8.2.4 Objective 4 – Data cleaning of responses to include spelling recommendations**

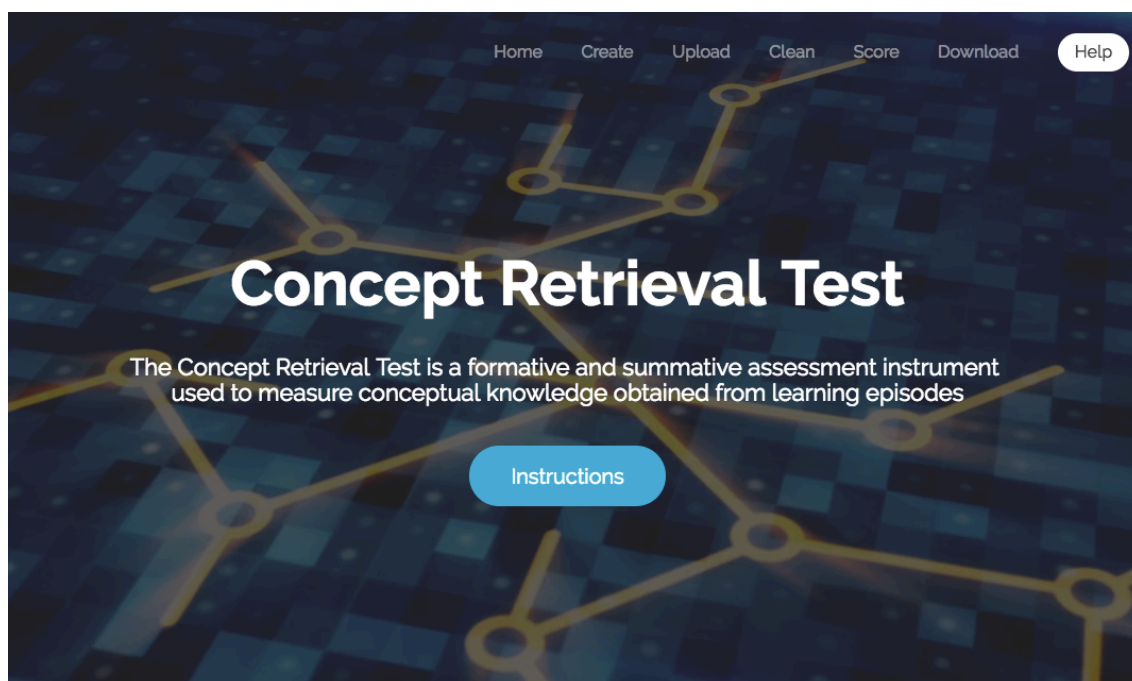
The previous version achieved the objective of utilising the spell-check submodule. However, the spelling error was only displayed to the user, not the spelling recommendation. This could cause a significant issue as the user does not know what change they are performing and



may reduce the inter-rater reliability. Therefore, this submodule will be enhanced to include the list of spelling mistakes, but also the spelling recommendations that the user will action. Finally, the opportunity to flag the used suffixes will also be explored to provide the user with more relevant information that can help inform the target word list.

### 8.2.5 Objective 5 – Improve user interface instructions and other help resources

The aim of this objective is to ensure that when the program is moved from the local to the online environment user errors are minimised. This can be achieved by a number of ways. Firstly, by providing users with appropriate instructions within the user interface prior to operating the instrument. Figure 8.1 is the homepage for the program and will require users to select the instructions command button prior to using the program. The instructions page will then help onboard new users to minimise errors. Once, users are familiar with the operation of the program they will utilise the navigation menu at the top of the page.



*Figure 8.1.* The homepage for the automated Concept Retrieval Technique (Version 3) including a button linking to the instructions page to provide users with appropriate instructions prior the using the program.

Secondly, the user interface is constructed to provide the user with appropriate opportunities for help and assistance by using label text. Figure 8.2 provides an example of the use of label text in the create database interface that assists the user in operating the program. They provide clear instructions regarding the type of data that is required from the user and further

information is provided that is formatted in bold. For example, in step 2 the information that is provided to user (i.e., Please select the number of concepts to be scored) provides a general instruction concerning the data to be inputted. The bold formatted label instruction provides more specific information (i.e., NOTE: This is how many concepts you have asked students to retrieve). This instruction links directly to the operation of the Concept Retrieval Technique to help the users if they are unclear in the operation. If there are any other issues not resolved, it would be expected that users would access the help pages and relevant technical documentation for creating a concept database. Finally, the user interface also needs to provide users with appropriate feedback when processes are executed (i.e., Create Database). As a result, displaying label text (i.e., Table created successfully) provides feedback to the user that they can move to the next process in operating the automated Concept Retrieval Technique.

The screenshot shows the 'Create Concept Database' interface. At the top is a dark navigation bar with the title 'Concept Retrieval Test' and links for Home, Create, Upload, Clean, Score, Download, and a Help button. The main heading is 'Create Concept Database'. Below it, 'Step 1: Setup Data Table' is shown with a red box around the instruction 'Please select the name of the primary key field for your CRT (NOTE: You only can have one primary key)' and a red arrow pointing to it. A text input field below contains 'Student name or ID'. 'Step 2: Number of Concepts' follows with a red box around the instruction 'Please select the number of concepts to be scored (NOTE: This is how many concepts you have asked students to retrieve)' and a red arrow pointing to it. A text input field below contains 'Max number of concepts'. At the bottom, a blue 'Create Database' button is shown, with a red box around the feedback message 'Table created successfully.' and a red arrow pointing to it.

Figure 8.2. The automated Concept Retrieval Technique interface highlighting label text that provides help information to users.

Thirdly, the help page is available to be accessed through the navigation menu at the top of the page, provides more extensive technical help for each process within the automated Concept Retrieval Technique. It is important that the technical help documentation follows some simple design rules, including removing technical language and ensuring every question that a user may ask is represented in the documentation with a possible answer. Furthermore, the documentation should be easy to navigate for the user to quickly locate the information and there should be a consistent use of screenshots to assist users navigate next steps. Finally, there are a number of third-party software applications that can be used to create the technical help documentation. However, given the limited number of processes in the automated Concept Retrieval Technique the documentation will be constructed manually by the programmer.

### **8.2.6 Objective 6 – Movement of the program from a local to an online environment**

In order to undertake beta testing the program will have to be moved from the local to an online environment. This will enable users to access the automated Concept Retrieval Technique from any geographical area via a URL. This will also ensure that user testing can be conducted to evaluate the stability of the program across different testing conditions and computer-based environments. The purchase of online storage is required specifically managing the MySQL databases. This will enable a unique domain to be generated and all program pages to be uploaded to the online domain.

## **8.3 System Modelling**

There are no changes to the system modelling tools that were utilized in version 1 and 2. This version will only employ some minor changes to the upload dataset submodule and clean dataset submodule. However, the download dataset submodule will be constructed to allow users the opportunity to export the scored results of the automated Concept Retrieval Technique (Version 3). The system flowcharts will help gain a deeper understanding to how each version objective is achieved in the implemented software solution.

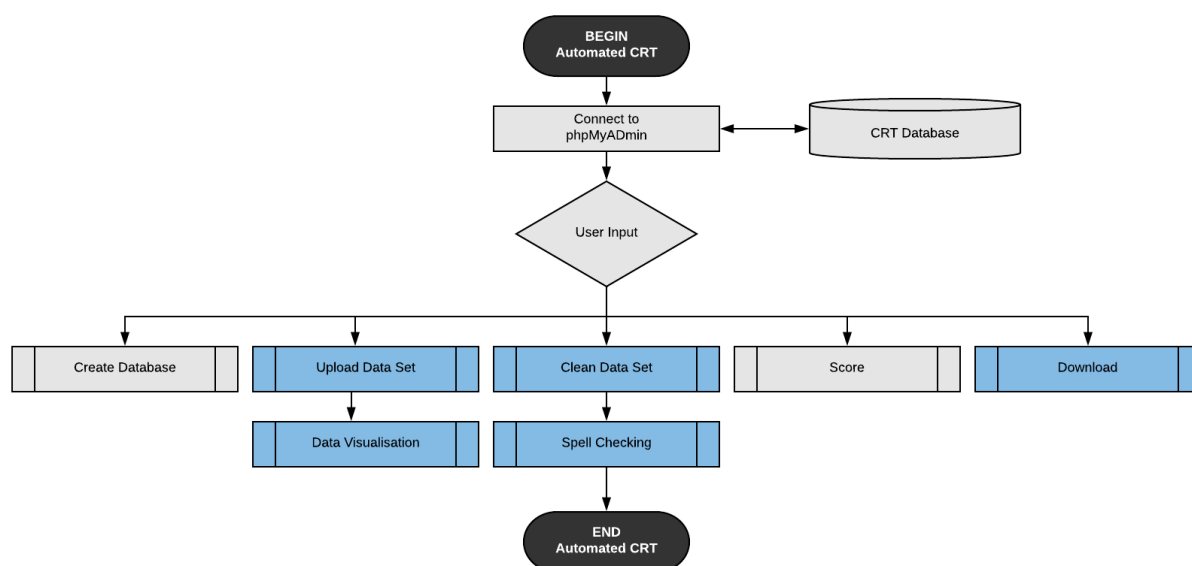
### **8.3.1 The automated Concept Retrieval Technique (Main Module)**

The main module was constructed in the automated Concept Retrieval Technique (Version 1), significant improvements were undertaken in the next version (Version 2) and some minor enhancements have been proposed with the current version (Version 3). Figure 8.3 represents the system flowchart for the main module with all submodules related to the current version isolated

in the colour blue. Firstly, the upload dataset submodule was enhanced in the previous version to include the calling of the data visualisation submodule. However, issues have been identified with redundant characters within test-taker responses that are causing run-time errors upon upload. Some minor changes will be required to this submodule to improve its stability, including feedback to the user when an unresolved error occurs. Furthermore, more user interactivity is required with the visualisation. Specifically, with the maximum number of concepts visualised and the colour schemes employed.

Secondly, the clean dataset submodule was constructed in the previous version. The aim of this submodule was to identify and display possible spelling issues to the user. The previous version was successful in identifying spelling issues, yet it was unable to display the spelling recommendations for the user to make a decision. Therefore, the current version will employ minor changes to achieve this objective. Given that minor changes will be made to both the upload dataset and clean dataset submodules there will not be any changes to the system flowcharts.

Finally, the download dataset submodule will be created to provide users the opportunity to download the scored results in csv format. This submodule will require establishing the connection to the online database and packaging the selected database table in preparation for download. In this version all fields will be exported including CRTID, responses and total scores. Once the submodule is executed the csv file will be created and downloaded to the user's desktop.



*Figure 8.3.* System flowchart of the automated Concept Retrieval Technique (Version 3) with the key submodules highlighted.

### 8.3.2 Download dataset (Submodule)

The primary aim of the download dataset submodule is to allow the user to download the scored Concept Retrieval Technique results addressing design objective 1. Figure 8.4 represents the system flowchart for the download dataset submodule. The first process is ensuring that a connection to the CRT MySQL database exists. Once a connection is made then the appropriate table can be identified for the dataset download. If a connection cannot be established then the program will flag an error and this will be displayed to the user. This could also occur if a connection is established, yet the appropriate table cannot be found. If no error is found then the download process will be executed, where all test-taker responses will be packaged in the appropriate format (i.e., CRTID, Concept1, Concept2, Score). This data will be stored temporarily as a text file until it is downloaded to the user and saved as a csv format. Finally, feedback will be provided to the user if the download was successful.

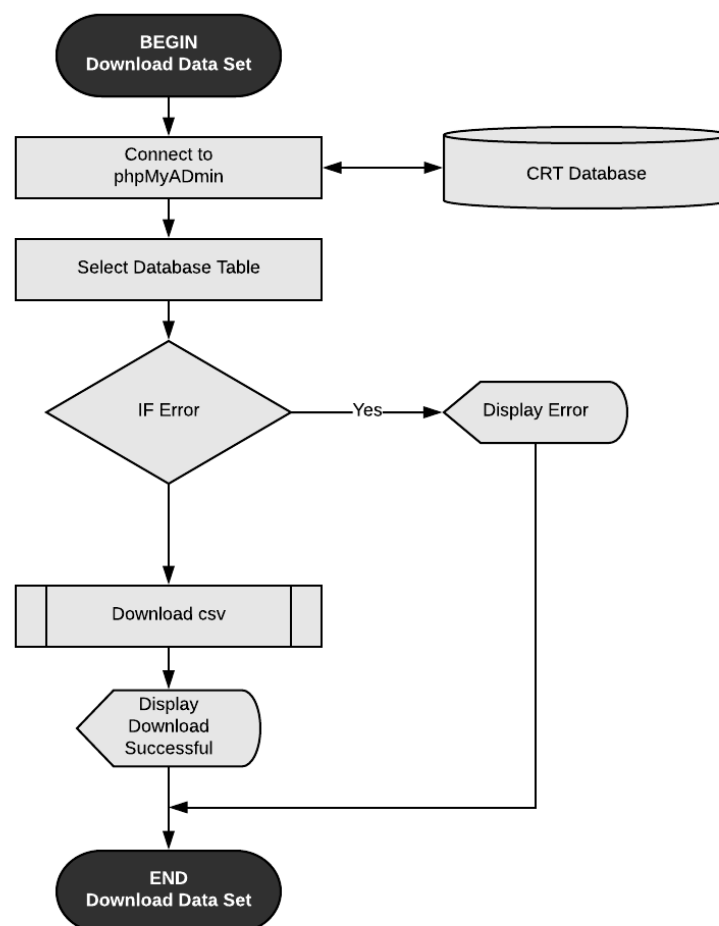


Figure 8.4. System flowchart of the download dataset submodule to including the use of a csv file.

## 8.4 Module Construction

The algorithms below have been provided in this thesis to help understand the programming logic and how the objectives have been achieved in the creation of the automated Concept Retrieval Technique (Version 3). Each of the algorithm flowchart provided will assist in providing a deeper understanding of the different processes, decision and iterations that are used to achieve the current version objectives.

### 8.4.1 Download dataset (Submodule)

The purpose of the download dataset submodule is to address design objective 1 of the automated Concept Retrieval Technique (Version 3). The final step of the automation process is the ability for the user to download the results of the test. It is important that these results are accessible to the user, firstly for the basis of changes to future instruction and secondly so that they can be uploaded into third party mark book or reporting programs so that they can be used for formal assessments. Figure 8.5 provides the algorithm representation for this submodule. The first process in this submodule is making a connection to the MySQL database and pointing to the current database table that had been created by the user. Next, the maximum number of rows (i.e., SMax) and the maximum number of columns (i.e., FMax) are extracted from the database table. Next, the program will start with the first test-taker and iterate through each field extracting the field data and storing it in a string with the comma character preceding each input (e.g., John13,proton,mass,). This iteration will continue until all the fields have been exhausted. The process will continue for the next test-taker in the database table and will continue until all test-takers have been exhausted. As a result, all data in the database table will be exported into a text file with commas separating each field. Finally, the text file will be exported as the csv format and downloaded onto the user's desktop for use with other programs. See Figure 8.6 for the interface for the download dataset submodule.

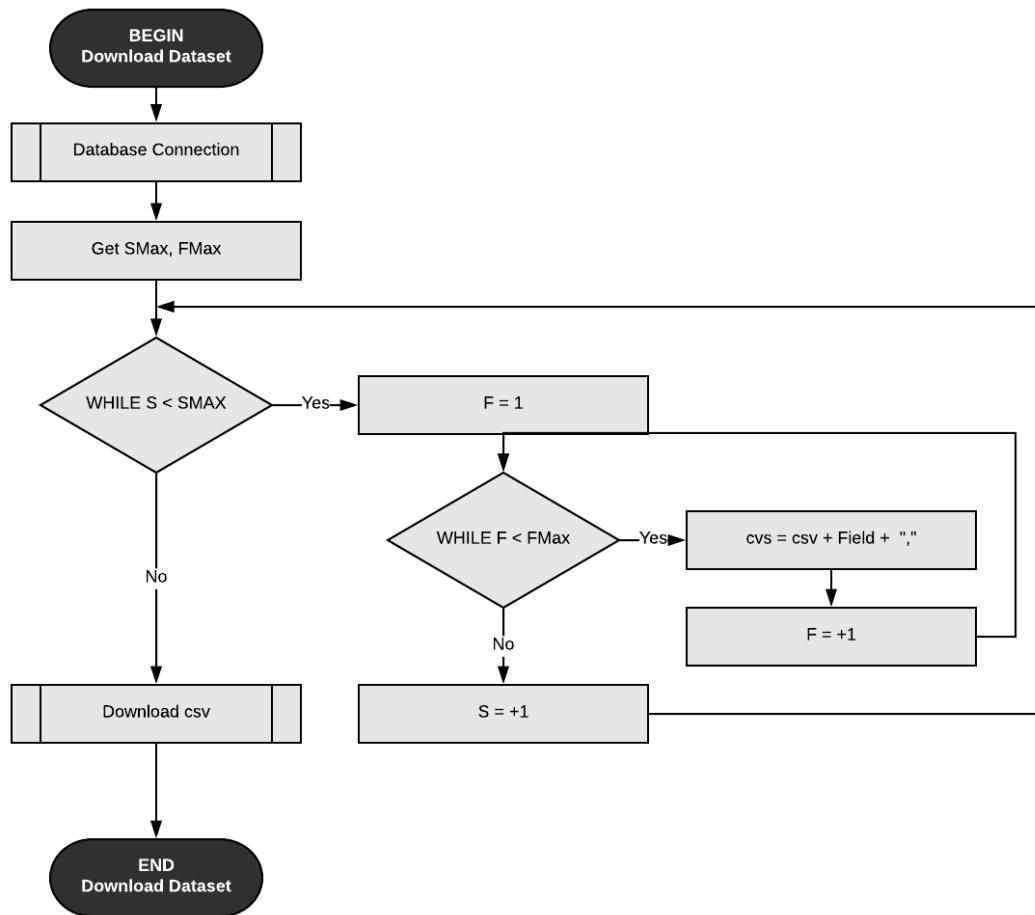


Figure 8.5. The download dataset submodule that downloads scored test-taker responses.

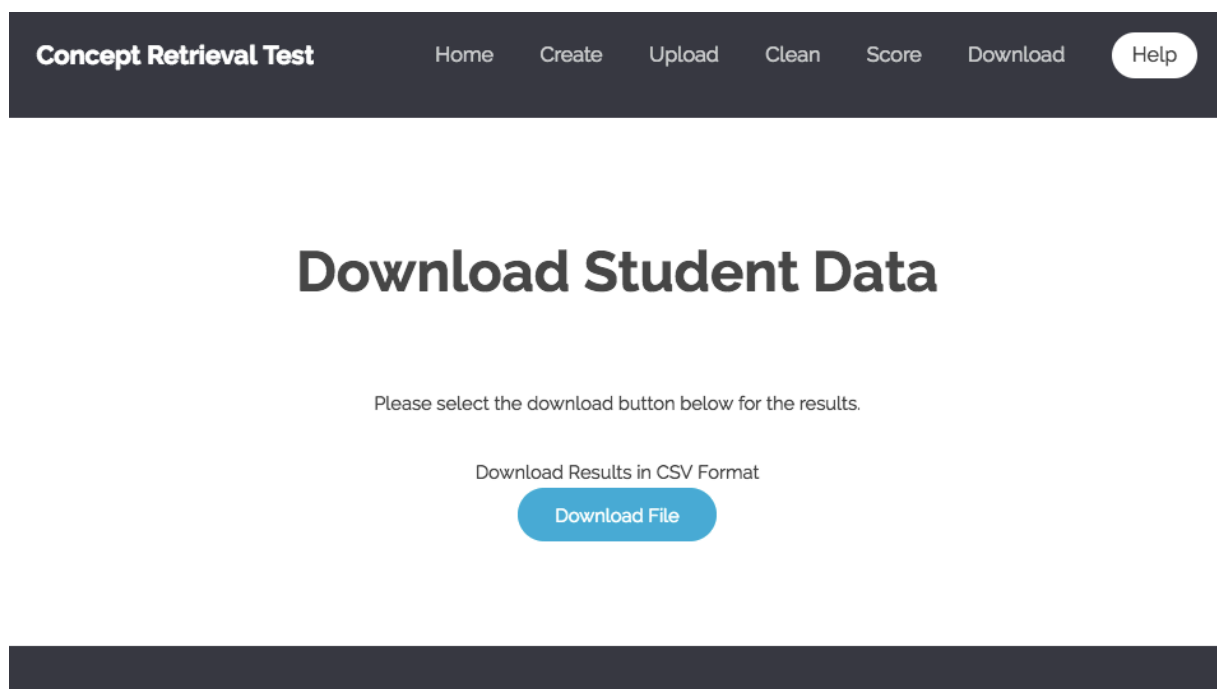


Figure 8.6. Interface for the download dataset submodule.

### 8.4.2 Upload dataset (Submodule)

The upload dataset submodule has been modified in the previous version to include the word cloud visualisation. However, the purpose of the modifications in the automated Concept Retrieval Technique (Version 3) is to address design objectives 2 and 3. The first design objective is focused on resolving issues with redundant characters on the uploading csv files to the online database. The most significant issue is the use of commas in test-taker responses (e.g., Protons, Neutrons and Electrons) as these characters are assumed in the input as a field separator to differentiate the different test-taker responses. Another issue is the use of the apostrophe in test-taker responses. Both issues generate a runtime error when using the upload dataset submodule and cause the program to crash. The ability to remove these characters from the upload dataset submodule will increase the functionality of the program.

The second objective focuses on providing some more interactivity for the user with the word cloud visualisation. Firstly, users have been provided the opportunity to enter via an input text box the maximum number of concepts to be visualised in the word cloud. Previously, this had been a fixed value of 15. However, for larger tests this value was considered as limiting and may not provide users with enough evidence from test-taker responses to inform the target word list. Secondly, a number of different colour scheme randomisations have been provided to the user in the interface. Figure 8.7 displays the interface for the upload dataset submodule with the four colour scheme options for the user.

## Upload Concepts

### Upload Student Data

Please select and upload your student data via csv file format.

Choose file No file chosen

### Visualisation Variables

Please select the maximum number of concepts to be visualised.

Please enter a number:

From the list box below please select the visualisation colour scheme

✓ Colour Scheme 1

Colour Scheme 2

Colour Scheme 3

Colour Scheme 4

Submit

### Concept Visualisation

Figure 8.7. Interface for the upload dataset submodule to include the changes to the visualisations.



### 8.4.3 Clean dataset (Submodule)

The clean dataset submodule was successfully constructed in the previous version. However, modifications are required to the automated Concept Retrieval Technique (Version 3) to address design objective 4. The primary issue is that the previous version was unable to display the spelling recommendations to the user. The achievement of this objective did not require any significant changes to the programming logic. Hence, new algorithm flowcharts were not required as the changes were only reflected in the programming code. See Figure 8.8 for the interface for the improved clean dataset submodule. The interface design for the current version employed the use of a table to handle and organise the data outputted, so that it is easy to interpret by the user.

Within the table the spelling issues is identified, a recommendation is provided and the user has the option to fix the issue. Once the command button (i.e., change) is selected any occurrence of the spelling issue in the online database will be changed. Finally, the opportunity to provide users with feedback regarding the use of suffixes was explored. However, the preliminary tests found it difficult to extract reliable data from test-taker responses that could be used to inform the target word list and have any positive effect on inter-rater reliability. Therefore, this feature of objective 4 was not achieved.

Spelling Issue Identified	Suggested Recommendation	Action
Alkalimetals	Alkalimeters	<button>Change</button>
Neautrons	Neutrons	<button>Change</button>
Magnesuim	Magnesium	<button>Change</button>
Valencey	Valency	<button>Change</button>
Reactivity	Reactivity	<button>Change</button>

Figure 8.8. Interface for the clean dataset submodule with the spelling recommendations included.

#### 8.4.4 Help (Submodule)

The help submodule will be constructed in the automated Concept Retrieval Technique (Version 3) to address design objective 5. There have been a number of improvements implemented to address this objective. Firstly, an instruction landing page for new users has been created that is accessed from the “instructions” command button on the home page. Figure 8.9 shows the interface for the instruction landing page, highlighting the three phases in operating the automated Concept Retrieval Technique (Version 3). The instructions provided on this page are basic, but help users understand that test-taker responses need to be uploaded, then cleaned and visualised, before they are scored and finally downloaded. These processes are replicated in the sequence of processes in the navigation menu bar at the top of each page. Users will then select the get started command button to begin the uploading of test-taker responses. Whereby, experienced users will not access the instruction landing page and will utilise the navigation bar to begin using the program.

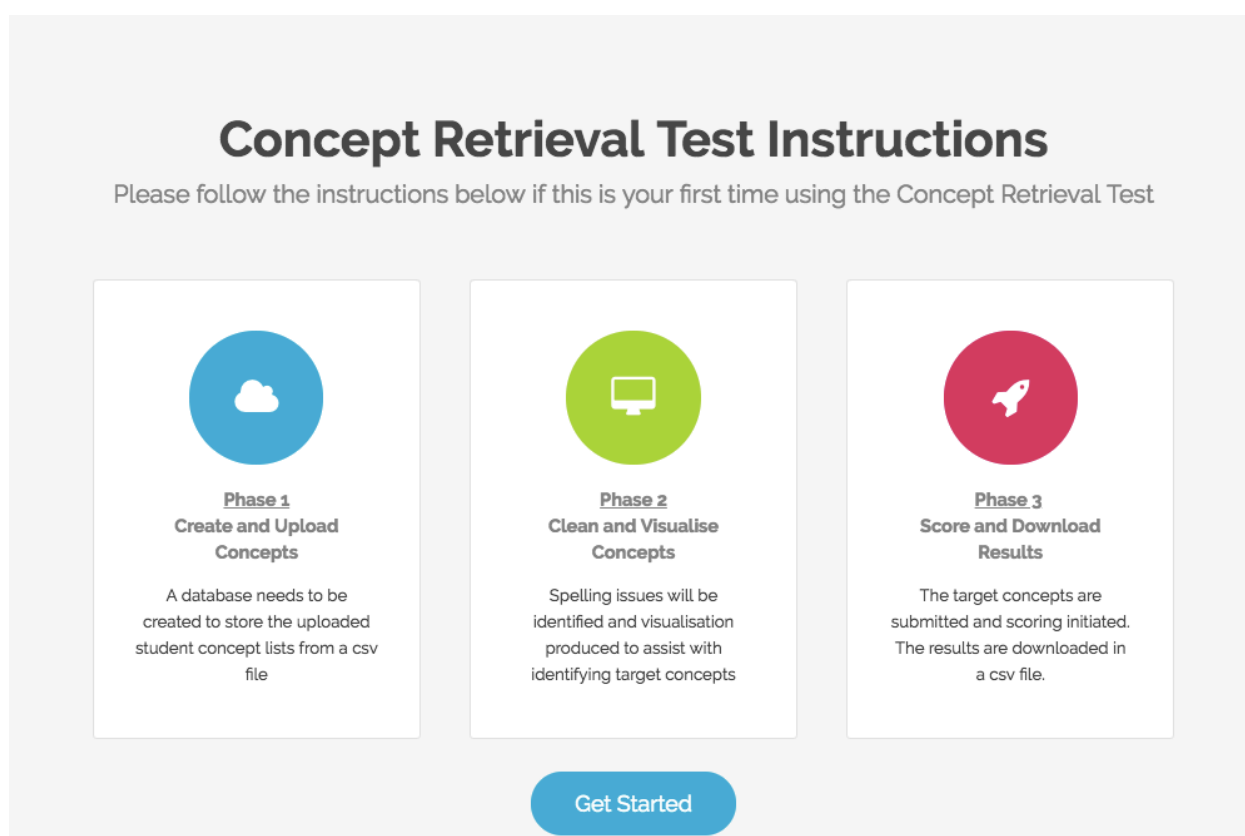


Figure 8.9. Interface for the instruction landing page to provide insights to new users.

Secondly, more specific and technical help documentation can be accessed via the help page. To access these resources users can select this page from the navigation menu located at the top of the page. This navigation button has been designed to be accessible from all pages and emphasis placed on the help button as shown in Figure 8.10. Furthermore, the documentation contained in this page should firstly provide tutorials for new users. These tutorials will be indexed according to the steps identified in operating the automated Concept Retrieval Technique. During alpha testing the commonly occurring issues were documented and the tutorials were created to alleviate these issues. Each tutorial includes a video with an expert explaining how to operate the program as shown in Figure 8.11. The other type of help provide is frequently asked questions (FAQ). These questions have been documented during beta testing and capture the commonly occurring questions asked by users in operating the test. Figure 8.12 shows that each question provides users with relevant test information and screenshots to help users navigate their own issues and avoid contacting the developer for minor issues.

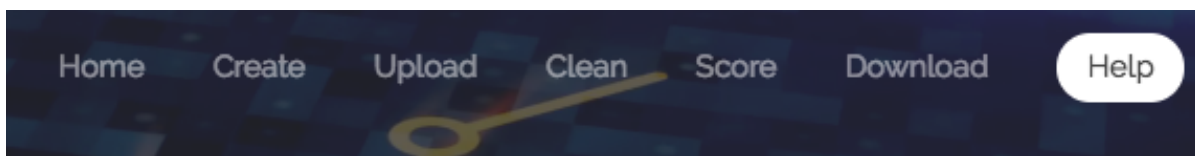
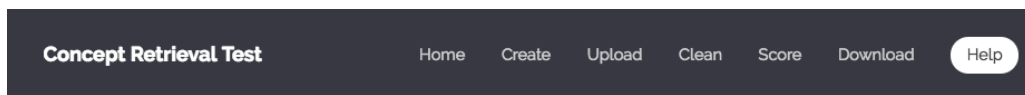


Figure 8.10. The navigation menu showing the emphasis on the help command button.



## Help Documentation

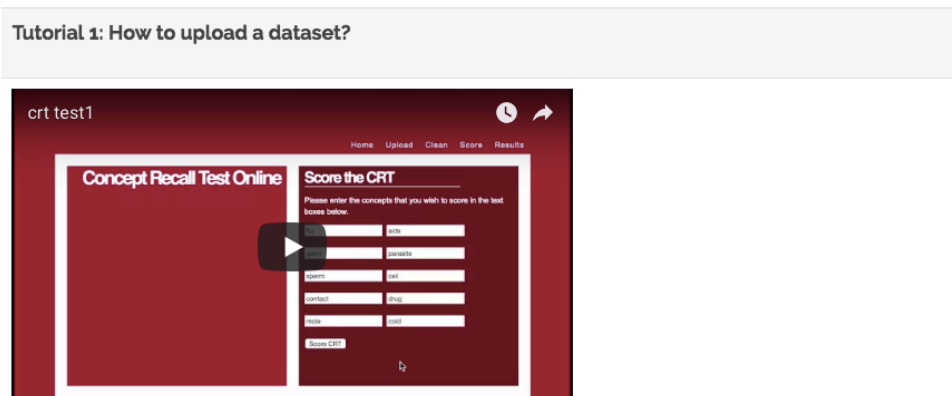


Figure 8.11. The interface for the help page showing an example of a tutorial to help users upload a dataset into the online database.

### Frequently Asked Questions (FAQ)

#### 1. How to change the layout of the word cloud visualisation?

When the visualisation has been created there are a number of variables that can be accessed by the user. The first (1) allows the user to change the number of concepts displayed in the visualisation. The default value is 15. However, users can increase this number according to the size of the CRT. The other variable is the option related to the different colour schemes (2) available to the word cloud. Users can select different visualisations and changes will be made to reflect these changes. Please see below a screenshot showing the different variables.

**Visualisation Variables**

Please select the maximum number of concepts to be visualised.

1 Please enter a number:

From the list box below please select the visualisation colour scheme

2 Colour Scheme 1

Submit

*Figure 8.12.* The interface for the Frequently Asked Questions page of the help menu showing the answer to a question including a screenshot example.

## 8.5 Online Environment for beta testing

It is planned that the automated Concept Retrieval Technique (Version 3) will be moved from the local environment to an online domain to enable end user beta testing. Suitable online server domains have been sourced. Once selected the MySQL database will be uploaded to the online server and the respective webpages will also be uploaded. The next step will be accessing the program from different testing sites and conditions and documenting its stability. There will be two phases for the online beta testing. The first will focus on functional testing, that is determining if the program functions as it is expected to perform. This type of testing will focus on the how users are able to navigate the automated Concept Retrieval Technique (Version 3) in the classroom environment. It is not expected that any significant errors will be identified. However, some minor errors may be discovered in the administration of the program and it is important that these minor version changes are provided via patches, to improve the overall effectiveness.

The second focus will be on performance testing. This involves measuring the performance of the servers. For example, the time taken to upload or score the test should be immediate. However, if there are issues with accessing the servers a time delay may occur. The performance of the servers need to be tested and documented to ensure appropriate feedback is given to users,

especially if small time delays do occur. Also, performance testing involves the targeting of different web browsers (e.g., Chrome and Safari) and how the program handles portable devices (e.g., iPad, Kindle, Surface). All the aspects of functional and performance testing are documented and used to construct the testing report provided to the programmer.

### **8.6 Study 9 – Stability of the Automated Concept Retrieval Technique (Version 3)**

In this study, a new set of data was used to explore the stability of the automated Concept Retrieval Technique (Version 3) administered over a significant period of time (i.e., eight iterations within a three-month period) and a larger number of test-takers (i.e.,  $n=175$ ). The aim of this current study was to measure the overall improvement of the current version with an increased “stress” placed on the instrument by the number of iterations and test-takers. The Concept Retrieval Technique was undertaken eight times in a period of three months for the secondary science topic of “chemical and physical changes”. Each test was administered to the same group of test-takers and two raters were used to determine the human inter-rater agreement for each iteration. Cohen’s kappa was used as the statistical measure to assess the agreement and correspondence between both raters. For each iteration, almost perfect agreement was observed by the human raters. Moreover, an agreed score was determined by the human raters and this score was used to measure the inter-rater reliability, stability and overall utility between machine to human scores for the automated Concept Retrieval Technique (Version 3). It is anticipated that this final study will provide significant evidence to demonstrate the overall suitability of the automated Concept Retrieval Technique (Version 3) as a measure of conceptual knowledge.

## **Method**

### **Participants**

On average one hundred and seventy-six secondary school science students participated in the study from an all-male secondary school in Australia. Their average age was 13 years. Participation was voluntary and students were not compensated for their participation.

### **Materials**

A Concept Retrieval Technique was developed for the topic of “chemical and physical changes”. A subject-matter expert worked together with the researchers to generate the target word list containing all admissible concepts for the topic. The target word lists used in administration of the Concept Retrieval Technique are shown in Table 8.1.

Table 8.1

*The target word list used for scoring the topic chemical and physical changes*

<u>Admissible concepts</u>	
Physical change	Gas formation
Chemical change	Precipitate
Temperature change	Respiration
Colour change	Photosynthesis
Properties	Weathering
New substance	Compound
Particle arrangement	Reverse
Pharmaceuticals	Polymer
Scientific research	Earth sphere
Reaction	

**Procedure**

The administration of the Concept Retrieval Technique was consistent with previous iterations. Whereby, test-takers were instructed to write down all concepts or ideas about the topic using only keywords or bullet points, with the directive to avoid writing in full sentences. Following the administration of the Concept Retrieval Technique, two raters independently marked the responses with the help of the target word list. One mark was awarded for each correctly identified concept and a final score for each test-taker was thereby generated. Table 8.2 shows the inter-rater agreement generated from the two raters for the Concept Retrieval Technique. The raters then resolved any inconsistencies in their scores to generate a final score. This score was used as a measure of inter-rater reliability against the scores generated from the automated Concept Retrieval Technique (Version 3). The operation of the automated Concept Retrieval Technique (Version 3) required the same list of student concepts to be imported into the online database. Furthermore, the same target word list used by human raters was imported and machine-scoring executed. Finally, the total scores of the machine-scoring process were downloaded and added to SPSS. Comparisons were made between both methods of assessment.

## Analysis

Inter-rater agreement was established by computing the Cohen's Kappa using the Statistical Package for the Social Sciences (SPSS), version 21.

## Results and Discussion

Overall, the administration of the automated Concept Retrieval Technique (Version 3) resulted in increased stability from previous versions and consistency in multiple iterations. The inter-rater reliability data presented in Table 8.2, demonstrates increased consistency between both the manual and automated scoring processes. This is a significant improvement in inter-rater reliability from version 1, where only slight agreement was observed and provides evidence for the overall stability and utility of the automated test.

Table 8.2

*The values of Cohen's  $\kappa$  calculated across the eight different studies for chemical and physical changes.*

<i>Sample</i>	<i>n</i>	<i>Manual (<math>\kappa</math>)</i>	<i>Automated (<math>\kappa</math>)</i>	<i>Agreement</i>
CRT1	177	.97	.95	Almost perfect
CRT2	174	.91	.94	Almost perfect
CRT3	177	.87	.96	Almost perfect
CRT4	177	.88	.97	Almost perfect
CRT5	178	.94	.95	Almost perfect
CRT6	175	.93	.95	Almost perfect
CRT7	172	.91	.91	Almost perfect
CRT8	172	.94	.95	Almost perfect

*Note:* The total number of students and kappa values for the reliability studies conducted across eight different studies for chemical and physical changes.

The average kappa obtained for the automated Concept Retrieval Technique (Version 3) across the eight tests was  $\kappa = .95$ , which demonstrates almost perfect agreement. Furthermore, the

increase in the number of test-takers within these studies also demonstrate that this test is highly reliable, especially across the spectrum of application in education (i.e., from class to grade level).

Given the improvements made in the current version and inter-rater reliability data collected, some issues still exist. Firstly, there are still some minor issues in the consistent spelling of target concepts within test-taker responses. Specifically, multisyllable words were identified as an issue throughout the administration of the eight iterations of the test. The occurrence of this issue primarily resided with the *Photosynthesis* and *Pharmaceuticals* concepts. In some instances, the spelling errors with these target concepts were so significant that the spell-checking submodule was unable to propose the correct word and subsequently they did not receive a mark in the machine-scoring process. Although this error was not frequent, it was a major contributor to the differences in inter-rater reliability between human and machine raters.

Secondly, the issue with the use of suffixes was not rectified in the current version and still remains an issue. Therefore, objective four was not achieved and there was no automated process to fix the issue with suffixes. Originally, the target word list contained two issues with suffixes that were rectified by consulting the subject matter expert and modifying the target concepts. The use of the concept *compounds* was the first issue as many test-takers had a mixture of responses including both versions of the concept (i.e., compound and compounds). The analysis of human rater results found that human raters had correctly marked both interpretations of the concept. Therefore, the target word list was modified to search and score *compound* as the target concept and this would also include other versions. The second issue occurred in the difference between *precipitate* and *precipitation*. Although, this issue only occurred once throughout the eight test administrations. This issue could have a significant effect on inter-rater reliability and may not be able to be rectified by modifying the target word list. A further elaboration regarding this issue will be provided in the next chapter.

## 8.7 Summary of key findings

In summary, the automated Concept Retrieval Technique (Version 3) provided substantial evidence of reliability of it in the educational setting. Overall, the average of the kappa was  $\kappa = .95$  across the eight administrations of the test, and a significant increase from the results obtained from the first version of the automated measure. There are still some areas that could be improved to ensure that the test can be used effectively in a classroom environment. Specifically, the construction of the target word list that is closest to the root word of each concept is an important attribute of future versions. Enhancements should include a range of methods to empower the user



with this understanding (i.e., user tutorials, instruction labels and dialog box prompts). It is important that ongoing beta testing is tracked and data obtained regarding how effective users are in constructing effective target word lists.

While this thesis only addresses inter-rater agreement as an indicator of the reliability of the Concept Retrieval Technique, one could also test-retest reliability as an important aspect of reliability especially with the automation. Test-retest reliability is defined as the extent to which the test results are consistent over time (Graham & Naglieri, 2003). To determine test-retest reliability, the test should be administered to the same participants on two or more separate occasions. This was achieved in Study 9 as the automated Concept Retrieval Technique was applied to eight administrations of test over a three-month period. As a result, observations supported the stability and consistency of inter-rater agreement.



# 9

Chapter

**Summary and Conclusions**

## 9.1 Introduction

This chapter concludes the proposal, validation and automation of a new measure of conceptual knowledge that is intended to be used within a variety of learning environments. This alternative to traditional assessment formats promotes the free recall of concepts or ideas for a given topic, rather than recognition. The primary aim of the thesis was to address the scoring reliability and validity of the test. Furthermore, evidence regarding how knowledge is organised and retrieved from the brain, was presented in support of the Concept Retrieval Technique procedure. Finally, the automated Concept Retrieval Technique was tested to explore its utility for education. In this chapter, the key findings will be reviewed and discussed as well as limitations and future research.

## 9.2 Summary of findings

The aim of the thesis was to present an alternative measure of conceptual knowledge. In chapter 1, the operationalization of the Concept Retrieval Technique was outlined as an instrument that requires test-takers to list all concepts and ideas that they can associate based on a trigger. The scoring process involves the use of a target word list to identify matches to test-taker responses. For every correctly recalled target concept the test-taker will receive one mark. Throughout the thesis the manual marking of the Concept Retrieval Technique involved two independent raters to assist in the analysis of interrater reliability. In addition, Chapter 1 introduced the research questions studied in this thesis. These questions were: (1) How reliable is the Concept Retrieval Technique as a new measure of conceptual knowledge to be used consistently across different school subjects and age groups? (2) Given the nature of the Concept Retrieval Technique, how can it be utilised as a valid measure of conceptual understanding? (3) What is the correlation between student performance on the Concept Retrieval Technique and their performance on the conventional assessments for the same topics? (4) What validity evidence can be provided to determine whether the Concept Retrieval Technique measures what it is intended to measure? and (5) How reliable is the machine-scoring of the Concept Retrieval Technique in comparison to human scoring?

In Chapter 2, a literature review on existing research which provided psychological and neuropsychological evidence to support the Concept Retrieval Technique and its application to education was presented. This literature supported the hypothesis that the Concept Retrieval Technique represents the knowledge network of an individual. As a result, based on spreading activation theory, the use of an input trigger will activate other concepts that are connected and

continue to spread along concept links until those links are exhausted. Thus, producing a schema of an individual's semantic network, which is based on well-established research findings from cognitive psychology on how knowledge is organized. The target word list is used to identify the correct concepts found in the schema and generate a total score that can be utilized by the teacher in the learning environment for differentiation or strategic student groupings.

In Chapter 3 and 4, five studies were undertaken to establish reliability and validity evidence for the Concept Retrieval Technique that requires the test-takers to freely recall relevant concepts about a specific topic. This was achieved and provided evidence that the construction of online automated marking process was necessary. The strengths of the Concept Retrieval Technique include its reliability and speed of application. Therefore, the use of machine-scoring will enable teachers to utilise real-time data to make immediate adjustments to their teaching within the lesson. Chapters 5, 6, 7 and 8 document the construction of the automated Concept Retrieval Technique. As opposed to human raters, the machine-scoring process is able to efficiently search and score a high frequency of test-taker responses. To support the construction of an automated Concept Retrieval Technique, a review of current evidence concerning automated assessment and machine-scoring was provided to examine the feasibility of constructing the online test. A range of considerations were explored such as an efficient method of importing test-taker responses, online data storage options and file formats to be utilized when exporting scored responses. There were three versions of the automated Concept Retrieval Technique, with each version incorporating enhancements based on the alpha testing of the previous version. Four studies were conducted to assess the inter-rater reliability, generalizability and overall utility of the automation given the previous evidence had been provided concerning the reliability and validity of manual raters. Below I will provide a summary of the findings, guided by the identified research questions.

### **1. How reliable is the Concept Retrieval Technique as a new measure of conceptual knowledge to be used consistently across different school subjects and age groups?**

In Study 1, the reliability of the Concept Retrieval Technique was examined by comparing the degree of agreement between two independent raters using a predetermined target word list of admissible concepts to score the students' responses. The results of this study revealed that the Concept Retrieval Technique was a reliable measure of conceptual knowledge demonstrated by the kappa being  $\kappa = .85$ , which suggested "almost perfect agreement". Next, Study 2 examined how consistent or generalizable the inter-rater agreement was with different raters and across a broad

spectrum of variables (i.e., subject domains and age groups). These results revealed across all conditions a high inter-rater agreement, with the kappa ranging from  $\kappa = .85$  to 1.00 and an average kappa of  $\kappa = .92$ . Overall, the consistently high inter-rater agreement supports the Concept Retrieval Technique as a reliable measure of conceptual knowledge. Finally, Study 3 was used to determine the reliability of the test in scoring concepts within full sentences. The same Concept Retrieval Technique was measured at three different opportunities to assess how reliable the measure was concerning its test-retest reliability. Overall, these findings support the effectiveness of the Concept Retrieval Technique as an appropriate measure of short answer responses, with a high test-retest reliability.

## **2. Given the nature of the Concept Retrieval Technique, how can it be utilised as a valid measure of conceptual understanding?**

What are the discriminating features of the Concept Retrieval Technique, that sets it apart from other measures of student learning? The first is that we have shown it to be a highly reliable measure of knowledge acquired by students, more reliable than most of its competitors. Concept mapping and open-ended questions, as used in the day-to-day practice of education, tend to be reliable only to a limited extent, making decisions about student performance sometimes a murky process. Although it is possible to construct MCQ examinations that are highly reliable, its implementation often requires time-consuming production of large numbers of items and subsequent extensive item analysis. Second, construction and administration of the Concept Retrieval Technique is simple and does not require much teacher time, with the exception of the construction of a target word list against which the responses of the students need to be scored. Administration of a Concept Retrieval Technique usually takes less than five minutes, enabling frequent testing without much disruption to the ongoing learning.

## **3. What is the correlation between student performance on the Concept Retrieval Technique and their performance on the conventional assessments for the same topics?**

In Study 4, the convergent validity of the Concept Retrieval Technique was investigated. This was done by first correlating the scores of the Concept Retrieval Technique with the scores of short-answer items on a conventional exam about the same topic (i.e., the periodic table). The results revealed a large positive correlation between both measures. This suggests adequate convergent validity of the Concept Retrieval Technique.

#### **4. What validity evidence can be provided to determine whether the Concept Retrieval Technique measures what it is intended to measure?**

In Study 5, the construct validity of the Concept Retrieval Technique has been investigated to answer the question above. Within this study we manipulated the amount of knowledge students could acquire about the topic “infectious disease”. A treatment and a control group were used in this experiment, with the control group receiving an unequal amount of knowledge about this topic. To that end, first, baseline knowledge of students was assessed using a Concept Retrieval Technique at the start of the experiment. Second, concept retrieval was measured after students were given the opportunity to participate in a problem discussion on the topic. Third, after the treatment group received a text about the microscopic world whereas the control group read a text about evolution, a third Concept Retrieval Technique was taken. The results from this experiment revealed only significant changes in the Concept Retrieval Technique scores for the treatment group during the learning phase when they were given the opportunity to acquire new relevant knowledge about the problem at hand. We interpreted these findings as supportive evidence for the construct validity of the Concept Retrieval Technique.

#### **5. How reliable is the machine-scoring of the Concept Retrieval Technique in comparison to human scoring?**

After the automated Concept Retrieval Technique had been designed and its psychometric properties determined, next its overall reliability in comparison to human scoring was determined. This was achieved by three studies that built on improvements of previous versions of the software. Within each study significant improvements in the inter-rater reliability were generated by the automated Concept Retrieval Technique. In Study 6, the topic “periodic table” was examined and produced a substantial kappa of  $\kappa = .85$  utilizing manual raters for scoring the test. An agreed score was determined between the manual raters and this score was used to determine the inter-rater reliability between human and machine-scoring by using the automated Concept Retrieval Technique (Version 1). The results were not as expected, demonstrated by the only slight agreement of the kappa of  $\kappa = .16$ . In Study 7, the improvements identified from Version 1 were applied with the automated Concept Retrieval Technique (Version 2) and a reliability analysis was performed again on the topic “periodic table” using the same data. The kappa between human and machine-scoring increased to  $\kappa = .70$ , demonstrating substantial agreement.

In Study 8 we explored the generalizability of the automated Concept Retrieval Technique (Version 2) by conducting a reliability analysis of three different subject topic areas. The utilization

of manual raters produced an average inter-rater reliability kappa of  $\kappa = .78$ . The same process was used previously but with manual raters and was compared to the machine-scored test. The results suggest that the average inter-rater reliability kappa for the automated Concept Retrieval Technique (Version 2) was  $\kappa = .71$ . Both results demonstrate substantial agreement and highlights the increased stability of the machine-scoring processes. Finally, Study 9 was constructed to measure the stability of the automated Concept Retrieval Technique (Version 3) over eight iterations involving a larger number of test-takers. Across the eight iterations of the automated Concept Retrieval Technique (Version 3) the average inter-rater reliability kappa was  $\kappa = .95$ . This demonstrated almost perfect agreement between human and machine-scorers and provide substantial evidence of the stability and overall educational utility of the automated Concept Retrieval Technique (Version 3).

### 9.3 Shortcomings

There are a few shortcomings of the Concept Retrieval Technique that need to be mentioned. These shortcomings pertain to the administration of the Concept Retrieval Technique and the automation marking processes. The first shortcoming is that the Concept Retrieval Technique is a reliable and objective measure of students' knowledge only when its administration is unannounced. Unfortunately, if students are aware that the Concept Retrieval Technique will be administered beforehand, it is highly likely that students will try to prepare for the test by rote memorizing any recent concepts that they have learnt in class. Consequently, the scores would not be a true representation of the actual semantic network of students.

The second shortcoming of the Concept Retrieval Technique is that students need to be carefully instructed of what is expected in the administration of the test. Clear instructions are required from the teacher regarding the responses that are expected in the Concept Retrieval Technique. Importantly, the expectation that they are to only write down the concepts as keywords or bullet points needs to be conveyed to students, so that they don't spend unnecessary time capturing every idea in detail. The width of the text box in Qualtrics® will provide some restrictions on the length of responses provided by students, but constant feedback needs to be provided to students regarding writing concepts as keywords or bullet points. Note, that this does not result in less reliable test scores, but could result in students not having the time to adequately retrieve their knowledge or may increase the demand on raters in identifying the correct concepts.

The third shortcoming is concerned with the automation of the Concept Retrieval Technique, whereby users need to be aware of the use of suffixes by test-takers in their responses. Although the word-cloud visualisation may provide some insights to the use of suffixes, it may

not identify all possible occurrences. Therefore, users need to ensure that all concepts selected for the target word list are the closest to the root word of the concept. For example, a user may have a target concept such as *Reactions* and considering the use of suffixes may choose to use *React* instead as the target concept. Given that, test-takers would have *Reaction*, *Reactions* and *Reactive* scored correctly. This would provide the opportunity for the automated assessment to mirror the decisions made by an expert rater. Therefore, changes have already been made to the instruction labels that provide users with assistance when undertaking the scoring process in the automated Concept Retrieval Technique (Version 3). Figure 9.1 shows the new version of the scoring engine interface with the instruction label in red. This label along with specific tutorials in the help page will provide enough assistance to the user to reduce the frequency of these issues.

**Concept Retrieval Test** Home Create Upload Clean Score Download Help

## Scoring Engine

**STEP 1: Number of concepts in the target list**

Number of concepts

**STEP 2: Concept target list**

Please ensure every concept is separated by a **comma**

**Note: Please ensure that all concepts in the target word list are root words and don't include a suffix**

Target concept list

Score CRT

*Figure 9.1.* The interface for the scoring engine that highlights for the user that concepts in the target word list needs to be “root” words and free of any suffix.

The fourth shortcoming is also concerned with the automated Concept Retrieval Technique. Specifically, the current software solution is only a prototype and the programming is quite rigid. As a result, there is not significant scope for the program to be agile and responsive to the needs of different users. For example, the database schema only allows users to select the number of concept fields in the administration and not any other additional identifier information



(i.e., class codes, sex or age). Any inputs that are made that do not conform to the program structures have the potential to cause a runtime error. Therefore, it is important that the services of a professional programmer are obtained to use the prototype to construct a commercial application. This solution will also include an opportunity to administer the test, score and download test-taker response data. A significant component of this thesis has been the design, implementation and reliability analysis of an automated version of the Concept Retrieval Technique. Therefore, the next step will be outsourcing of a programmer and the development of a professional software application that is built upon the work undertaken in this thesis.


#### **9.4 Directions for further research**

The findings of this thesis provide unique opportunities to improve the effective use of the Concept Retrieval Technique in an educational setting. There is need, however, to follow-up these studies with two further investigations into the administration of the Concept Retrieval Technique. Firstly, a method by which teachers can derive a target word list from learning objectives or lesson materials is paramount. Currently, word cloud generators can be used to identify frequently used concepts within a text or learning materials, however, a more precise and reliable methodology is required to ensure that a target word list is reflective of the key concepts and their connections as covered within a lesson or text, rather than simply showing the frequency of the use of that term. There exists the possibility that this process could be automated in the future via the creation of a specific application (especially if a text alone precedes the administration of the Concept Retrieval Technique). Moreover, a simple scripted procedure by which teachers administer the lesson (or part thereof) that is the subject of the Concept Retrieval Technique may also be required; especially if a text is to be accompanied by a tutorial involving direct instruction or elaboration on the topic in focus. This is to ensure the key concepts are not diluted or superseded by the instruction or responses of the teacher.

On a different note, some have criticized the approach as too simple. If learning is complex and cognition is complex, how can assessment be so simple? Our answer is that the ability to retrieve concepts from memory is the simple by-product of admittedly complex processes. Processing information by a computer may be complex, but retrieving information from a resulting database is (usually) simple. Others point to the fact that education ideally would instil in students higher-order cognitive abilities, such as the ability to solve appropriate problems, not simply knowledge. While we concur with this point of view, we would like to stress that problem-solving, like other higher-order cognitive processes, such as decision making and judgement, is entirely

knowledge-based; without appropriate knowledge no problem can be solved, as studies in human high-level expertise have demonstrated time and again (K.A. Ericsson, Charness, Feltovich, & Hoffman, 2006). (This is of course not to say that having appropriate knowledge is a *sufficient* condition for expertise to emerge. Beyond acquiring it students have to *apply* their knowledge in problem-solving situations. But this requires practice-oriented training and related assessment). Third, one could argue that not all concepts are equally important and that the Concept Retrieval Technique does not adequately represent this state of affairs. This objection can be easily countered by suggesting that teachers should indicate which concepts in their target word list should receive a higher weight and compute final scores taking into account that the presence or absence of these concepts should be taken into account. Based on our experience we predict however that weighing responses will not make much of a difference in the total scores of students.





# **Samenvatting** **(Summary in Dutch)**

## Introductie

Dit hoofdstuk sluit het voorstel, de validatie en automatisering van een nieuwe maat van conceptuele kennis die bedoeld is om te worden gebruikt in een verscheidenheid van leeromgevingen. Dit alternatief voor traditionele beoordelingsformaten bevordert het vrij terugroepen van concepten of ideeën voor een bepaald onderwerp in plaats van erkenning. Het primaire doel van het proefschrift was om de scoring betrouwbaarheid en validiteit van de test te valideren. Verder werd bewijs gepresenteerd over hoe kennis wordt georganiseerd en uit het geheugen wordt terug gehaald, ter ondersteuning van de Concept Retrieval Technique-procedure. Ten slotte werd de geautomatiseerde Concept Retrieval Technique getest om de bruikbaarheid ervan voor het onderwijs te onderzoeken. In dit hoofdstuk worden de belangrijkste bevindingen herzien en besproken, evenals beperkingen en toekomstig onderzoek.

## Samenvatting van bevindingen

Het doel van het proefschrift was om een alternatieve manier van het meten voor conceptuele kennis te presenteren. In hoofdstuk 1 werd de operationalisering van de Concept Retrieval Technique geschetst als een instrument dat van testpersonen verlangt dat ze alle concepten en ideeën opsomden die ze kunnen associëren op basis van een trigger. Bij het scoringsproces wordt een lijst met doelwoorden gebruikt om overeenkomsten met de antwoorden van de testpersoon te identificeren. Voor elk correct opgeroepen doelconcept krijgt de testpersoon een(1) punt. Gedurende het hele proefschrift omvatte de handmatige markering van de Concept Retrieval Technique twee onafhankelijke beoordelaars om te helpen bij de analyse van de interbeoordelaars betrouwbaarheid. Daarnaast introduceerde hoofdstuk 1 de onderzoeksvragen die in dit proefschrift zijn bestudeerd. Deze vragen waren: (1) Hoe betrouwbaar is de Concept Retrieval Technique als een nieuwe manier van het meten voor conceptuele kennis die consequent over verschillende schoolvakken en leeftijdsgroepen kan worden gebruikt? (2) Gezien de aard van de Concept Retrieval Technique, hoe kan het worden gebruikt als een geldige maatstaf voor conceptueel begrip? (3) Wat is het verband tussen de prestaties van studenten op de Concept Retrieval Technique en hun prestaties op de conventionele beoordelingen voor dezelfde onderwerpen? (4) Welke validiteitsbewijzen kunnen worden verstrekt om te bepalen of de Concept Retrieval Technique meet wat het is bedoeld om te meten? en (5) Hoe betrouwbaar is de scoringstechniek van de Concept Retrieval Technique in vergelijking met menselijke scores?

In hoofdstuk 2 werd een literatuuroverzicht gepresenteerd van bestaand onderzoek dat psychologisch en neuropsychologisch bewijsmateriaal ter ondersteuning van de Concept Retrieval Technique en de toepassing ervan op het onderwijs leverde. Deze literatuur ondersteunde de hypothese dat de Concept Retrieval Technique het kennisnetwerk van een individu vertegenwoordigt. Als gevolg hiervan, op basis van de spreidingsactivatietheorie, zal het gebruik van een ingangstrigger andere concepten activeren die zijn verbonden en zich blijven verspreiden langs conceptlinks totdat die koppelingen tot het einde zijn gekomen. Dus, het produceren van een schema van het semantische netwerk van een individu, dat is gebaseerd op gevestigde onderzoeksresultaten uit de cognitieve psychologie over hoe kennis is georganiseerd. De lijst met doelwoorden wordt gebruikt om de juiste concepten in het schema te identificeren en een totale score te genereren die door de docent in de leeromgeving kan worden gebruikt voor differentiatie of strategische studentengroeperingen.

In hoofdstuk 3 en 4 werden vijf studies uitgevoerd om de betrouwbaarheid en validiteit van de Concept Retrieval Technique vast te stellen, waarbij de testpersonen vrij moeten denken aan relevante concepten over een specifiek onderwerp. Dit werd bereikt en gaf het bewijs dat de constructie van een online geautomatiseerd markeringsproces noodzakelijk was. De waarde van de Concept Retrieval Technique omvatten de betrouwbaarheid en snelheid van toepassing. Daarom maakt het gebruik van scoring op de machine het leraren mogelijk om real-time gegevens te gebruiken en om onmiddellijk hun lessen in de klas aan te passen. De hoofdstukken 5, 6, 7 en 8 documenteren de constructie van de geautomatiseerde Concept Retrieval Technique. In tegenstelling tot menselijke beoordelaars, is het scoreproces in de machine in staat om efficiënt een hoge frequentie van antwoorden van testafnemers te zoeken en te scoren. Ter ondersteuning van de constructie van een geautomatiseerde Concept Retrieval Technique, werd een overzicht van de huidige gegevens met betrekking tot geautomatiseerde beoordeling en machinescoren verstrekt om de mogelijkheid van het samenstellen van de online test te onderzoeken. Een reeks van overwegingen werd onderzocht, zoals een efficiënte methode voor het importeren van antwoorden van testafnemers, online 'data storage' opties en "file" formaten die moeten worden gebruikt bij het exporteren van antwoorden met een score. Er waren drie versies van de geautomatiseerde Concept Retrieval Technique, waarbij elke versie verbeteringen bevatte, gebaseerd op de alpha-test van de vorige versie. Vier studies werden uitgevoerd om de interbeoordelaars betrouwbaarheid, de generaliseerbaarheid en de algemene nuttigheid van de automatisering te beoordelen, gezien het eerdere bewijsmateriaal met betrekking tot de

betrouwbaarheid en validiteit van handmatige beoordelaars. Hieronder geef ik een samenvatting van de bevindingen, geleid door de geïdentificeerde onderzoeksvragen.

### **1. Hoe betrouwbaar is de Concept Retrieval Technique als een nieuwe maatstaf voor conceptuele kennis die consistent moet worden gebruikt voor verschillende schoolvakken en leeftijdsgroepen?**

In Studie 1 werd de betrouwbaarheid van de Concept Retrieval Technique onderzocht door de mate van overeenstemming tussen twee onafhankelijke beoordelaars te vergelijken met behulp van een vooraf bepaalde doelwoordenlijst van toelaatbare concepten om de antwoorden van de studenten te scoren. De resultaten van deze studie onthulden dat de Concept Retrieval Technique een betrouwbare maatstaf was voor conceptuele kennis, aangetoond door de Kappa die  $K = .85$  was, wat een "bijna perfecte overeenkomst" suggereerde. Vervolgens onderzocht Studie 2 hoe consistent of generaliseerbaar de interbeoordelaar overeenkomst was met verschillende beoordelaars en over een breed spectrum van variabelen (d.w.z. vakdomeinen en leeftijdsgroepen). Deze resultaten onthulden in alle omstandigheden een hoge interbeoordelaar overeenkomst, waarbij de kappa varieerde van  $\kappa = .85$  tot 1.00 en een gemiddelde kappa van  $\kappa = .92$ . Over het algemeen ondersteunt de consistent hoge interbeoordelaar overeenkomst de Concept Retrieval Technique als een betrouwbare maatstaf voor conceptuele kennis. Ten slotte werd met Studie 3 de betrouwbaarheid van de test in scoreconcepten binnen volledige zinnen bepaald. Dezelfde Concept Retrieval Technique werd gemeten op drie verschillende manieren om te beoordelen hoe betrouwbaar de meting was met betrekking tot de test-hertest betrouwbaarheid. Al met al ondersteunen deze bevinding de effectiviteit van de Concept Retrieval Technique als een geschikte maatstaf voor korte antwoorden, met een hoge test-hertest betrouwbaarheid.

### **2. Gezien de aard van de Concept Retrieval Technique, hoe kan het worden gebruikt als een geldige maatstaf voor conceptueel begrip?**

Wat zijn de onderscheidende kenmerken van de Concept Retrieval Technique, die het onderscheidt van andere maatregelen voor het leren van studenten? De eerste is dat we hebben aangetoond dat het een zeer betrouwbare maatstaf is voor kennis die door studenten is opgedaan, betrouwbaarder dan de meeste van zijn concurrenten. Conceptmapping en open vragen, zoals gebruikt in de dagelijkse praktijk van het onderwijs, zijn over het algemeen slechts in beperkte

mate betrouwbaar, waardoor beslissingen over de leerprestaties soms een *duister proces* is. Hoewel het mogelijk is om MCQ-onderzoeken te construeren die zeer betrouwbaar zijn, de implementatie ervan vereist vaak een tijdrovende productie van grote aantallen items en daaropvolgende uitgebreide artikelanalyse. Ten tweede, de constructie en het beheer van de Concept Retrieval Technique is eenvoudig en vereist niet veel tijd van de leraar, met uitzondering van de constructie van een doelwoordenlijst waartegen de antwoorden van de studenten gescoord moeten worden. Het beheer van een Concept Retrieval Technique duurt meestal minder dan vijf minuten, waardoor frequent testen mogelijk is zonder veel verstoring van het voortgaande leren.

### **3. Wat is de correlatie tussen de prestaties van studenten op de Concept Retrieval Technique en hun prestaties op de conventionele beoordelingen voor dezelfde onderwerpen?**

In Studie 4 werd de convergente validiteit van de Concept Retrieval Technique onderzocht. Dit werd gedaan door eerst de scores van de Concept Retrieval Technique te correleren met de scores van kort-antwoord items op een conventioneel onderzoek over hetzelfde onderwerp (d.w.z. het periodiek tafel). De resultaten toonden een grote positieve correlatie tussen beide maatregelen. Dit suggereert een voldoende convergente validiteit van de Concept Retrieval Technique.

### **4. Welke validiteits bewijzen kunnen worden verstrekt om te bepalen of de Concept Retrieval Technique meet, wat het is bedoeld om te meten?**

In Studie 5 is de constructvaliditeit van de Concept Retrieval Technique onderzocht om de bovenstaande vraag te beantwoorden. Binnen deze studie manipuleerden we de hoeveelheid kennis die studenten konden verwerven over het onderwerp "infectieziekte". Een behandeling en een controlegroep werden in dit experiment gebruikt, waarbij de controlegroep een ongelijke hoeveelheid kennis over dit onderwerp ontving. Daartoe werd eerst basiskennis van studenten beoordeeld aan de hand van een Concept Retrieval Technique aan het begin van het experiment. Ten tweede werd concept retrieval gemeten nadat studenten de kans hadden gekregen om deel te nemen aan een probleem discussie over het onderwerp. Ten derde, nadat de behandelingsgroep een tekst ontving over de microscopische wereld terwijl de controlegroep een tekst las over evolutie, werd een derde Concept Retrieval Technique gevolgd. De resultaten van dit experiment toonden alleen significante veranderingen in de Concept Retrieval Technique-scores voor de



behandelingsgroep tijdens de leerfase, toen ze de kans kregen om nieuwe relevante kennis over het probleem bij de hand te krijgen. We interpreteerden deze bevindingen als ondersteunend bewijs voor de constructvaliditeit van de Concept Retrieval Technique.

### **5. Hoe betrouwbaar is de scoring van de Concept Retrieval Technique in vergelijking met het scoren van mensen?**

Nadat de geautomatiseerde Concept Retrieval Technique was ontworpen en de psychometrische eigenschappen ervan waren bepaald, werd vervolgens de algehele betrouwbaarheid in vergelijking met menselijke scores bepaald. Dit werd bereikt door drie onderzoeken die voortbouwden op verbeteringen van eerdere versies van het software. Binnen elk onderzoek werden significante verbeteringen in de interbeoordelaars betrouwbaarheid gegenereerd door de geautomatiseerde Concept Retrieval Technique. In Studie 6 werd het onderwerp "periodiek tafel" onderzocht en produceerde een substantiële kappa van  $\kappa = .85$  met behulp van handmatige beoordelaars voor het scoren van de test. Een overeengekomen score werd bepaald tussen de handmatige beoordelaars en deze score werd gebruikt om de interbeoordelaars betrouwbaarheid tussen mens en machine scoren te bepalen met behulp van de geautomatiseerde Concept Retrieval Technique (versie 1). De resultaten waren niet zoals verwacht, aangetoond door de enige lichte overeenstemming van de kappa van  $\kappa = .16$ . In Studie 7 werden de verbeteringen uit Versie 1 toegepast met de geautomatiseerde Concept Retrieval Technique (Versie 2) en een betrouwbaarheidanalyse werd opnieuw uitgevoerd met betrekking tot het onderwerp "periodieke tafel" met dezelfde gegevens. De kappa tussen het scoren van mensen en machines nam toe tot  $\kappa = .70$ , wat een aanzienlijke mate van overeenstemming aantoonde.

In Studie 8 hebben we de generaliseerbaarheid van de geautomatiseerde Concept Retrieval Technique (Versie 2) onderzocht door een betrouwbaarheidsanalyse uit te voeren van drie verschillende onderwerpsonderwerpen. Het gebruik van handmatige beoordelaars produceerde een gemiddelde interbeoordelaars betrouwbaarheids kappa van  $\kappa = .78$ . Hetzelfde proces werd eerder gebruikt, maar met handmatige beoordelingen en werd vergeleken met de machine-gescorede test. De resultaten suggereren dat de gemiddelde interbeoordelaars betrouwbaarheids kappa voor de geautomatiseerde Concept Retrieval Technique (versie 2)  $\kappa = .71$  was. Beide resultaten laten een grote mate van overeenstemming zien en benadrukken de toegenomen stabiliteit van de machine-scoringsprocessen. Ten slotte is Studie 9 geconstrueerd om de stabiliteit van de geautomatiseerde Concept Retrieval Technique (Versie 3) te meten over acht iteraties met

een groter aantal testpersonen. Over de acht iteraties van de geautomatiseerde Concept Retrieval Technique (versie 3) was de gemiddelde interbeoordelaars betrouwbaarheids  $\kappa = .95$ . Dit toonde een bijna perfecte overeenkomst tussen mens en machine-scoorders en leverde substantieel bewijs van de stabiliteit en het algehele educatieve nut van de geautomatiseerde Concept Retrieval Technique (versie 3).

### **Tekortkomingen**

Er zijn een paar tekortkomingen van de Concept Retrieval Technique die moeten worden genoemd. Deze tekortkomingen hebben betrekking op het beheer van de Concept Retrieval Technique en de automatiserings markerings processen. De eerste tekortkoming is dat de Concept Retrieval Technique alleen een betrouwbare en objectieve maatstaf is voor de kennis van studenten wanneer de administratie onaangekondigd is. Jammer genoeg, als studenten zich ervan bewust zijn dat de Concept Retrieval Technique van tevoren wordt toegediend, is het zeer waarschijnlijk dat studenten zich zullen proberen voor te bereiden op de test door uit de recente concepten te leren die ze in de les hebben geleerd. Bijgevolg zouden de scores geen echte weergave zijn van het daadwerkelijke semantische netwerk van studenten.

De tweede tekortkoming van de Concept Retrieval Technique is dat studenten zorgvuldig geïnstrueerd moeten worden over wat er verwacht wordt bij het afnemen van de test. Van de docent zijn duidelijke instructies nodig met betrekking tot de antwoorden die worden verwacht in de Concept Retrieval Technique. Belangrijk is dat de verwachting dat ze de concepten alleen als sleutelwoorden of opsommingspunten moeten neerschrijven, moet worden overgebracht naar studenten, zodat ze geen onnodige tijd besteden aan het vastleggen van elk idee in detail. De breedte van het tekstvak in Qualtrics® geeft enige beperkingen aan de lengte van antwoorden van studenten, maar constante feedback moet aan studenten worden verstrekt over het schrijven van concepten als sleutelwoorden of opsommingsstroken. Merk op dat dit niet leidt tot minder betrouwbare testcores, maar ertoe kan leiden dat studenten niet de tijd hebben om hun kennis adequaat op te halen of dat ze de vraag naar beoordelaars kunnen verhogen bij het identificeren van de juiste concepten.

De derde tekortkoming betreft de automatisering van de Concept Retrieval Technique, waarbij gebruikers zich bewust moeten zijn van het gebruik van achtervoegsels door testpersonen in hun antwoorden. Hoewel de visualisatie van word-cloud een aantal inzichten biedt in het gebruik van achtervoegsels, het is mogelijk niet alle mogelijke occurrences worden geïdentificeerd. Daarom moeten gebruikers ervoor zorgen dat alle concepten die voor de doelwoordenlijst zijn

geselecteerd het dichtst bij het hoofdwoord van het concept liggen. Een gebruiker kan bijvoorbeeld een doelconcept hebben zoals reacties en gezien het gebruik van achtervoegsels kan ervoor gekozen worden om React in plaats daarvan als doelconcept te gebruiken. Gezien dit feit testers zouden Reaction, Reactions en Reactive correct gescoord hebben. Dit zou de mogelijkheid bieden voor de geautomatiseerde beoordeling om de beslissingen van een expert-beoordelaar te weerspiegelen. Daarom zijn er al wijzigingen aangebracht in de instructielabels die gebruikers helpen bij het uitvoeren van het scoreproces in de geautomatiseerde Concept Retrieval Technique (versie 3). Afbeelding 9.1 toont de nieuwe versie van de scoring-engine-interface met het instructielabel in rood. Dit label en specifieke zelfstudies op de helppagina bieden voldoende ondersteuning aan de gebruiker om de frequentie van deze problemen te verminderen.

**Concept Retrieval Test** Home Create Upload Clean Score Download Help

## Scoring Engine

**STEP 1: Number of concepts in the target list**

Number of concepts

**STEP 2: Concept target list**

Please ensure every concept is separated by a **comma**

**Note: Please ensure that all concepts in the target word list are root words and don't include a suffix**

Target concept list

Score CRT

*Figuur 9.1.* De interface voor de scorende engine die voor de gebruiker benadrukt dat concepten in de lijst met doelwoorden "root" woorden moeten zijn en vrij van elk achtervoegsel.

De vierde tekortkoming betreft ook de geautomatiseerde Concept Retrieval Technique. Concreet is dat de huidige software-oplossing slechts een prototype is en de programmering vrij 'rigid' is. Als gevolg hiervan is er geen significante ruimte voor het programma om behendig te zijn en te reageren op de behoeften van verschillende gebruikers. Het databaseschema staat

bijvoorbeeld alleen gebruikers toe om het aantal conceptvelden in de administratie te selecteren en niet enige andere aanvullende identificatie-informatie (d.w.z., klassencodes, geslacht of leeftijd). Alle ‘input’ die zijn gemaakt die niet voldoen aan de programmastructuren, hebben de potentie om een runtime-fout te veroorzaken. Daarom is het belangrijk dat de diensten van een professionele programmeur worden verkregen om het prototype te gebruiken om een commerciële toepassing te bouwen. Deze oplossing biedt ook de mogelijkheid om de testgegevens te beheren, te scoren en te downloaden. Een belangrijk onderdeel van dit proefschrift was de ontwerp-, implementatie- en betrouwbaarheidsanalyse van een geautomatiseerde versie van de Concept Retrieval Technique. Daarom is de volgende stap het uitbesteden van een programmeur en de ontwikkeling van een professionele software toepassing die is gebaseerd op het werk dat in dit proefschrift is gedaan.

### **Mogelijkheden voor verder onderzoek**

De bevindingen van dit proefschrift bieden unieke kansen om het effectieve gebruik van de Concept Retrieval Technique in een educatieve setting te verbeteren. Er is echter behoefte aan follow-up van deze studies met twee verdere onderzoeken naar het beheer van de Concept Retrieval Technique. Ten eerste is een methode waarmee leraren een doelwoordenlijst kunnen afleiden uit leerdoelen of lesmateriaal van het grootste belang. Momenteel kunnen woordwolkgeneratoren worden gebruikt om vaak gebruikte concepten binnen een tekst of leermateriaal te identificeren, maar een meer precieze en betrouwbare methodologie is vereist om ervoor te zorgen dat een lijst met doelwoorden een afspiegeling is van de belangrijkste concepten en hun verbindingen zoals behandeld in een les of tekst, in plaats van simpelweg de frequentie van het gebruik van die term te laten zien. Er bestaat de mogelijkheid dat dit proces in de toekomst geautomatiseerd kan worden via de creatie van een specifieke applicatie (vooral als alleen een tekst voorafgaat aan het beheer van de Concept Retrieval Technique). Bovendien kan een eenvoudige procedure met scripts waarmee leraren de les (of een deel daarvan) die het onderwerp is van de Concept Retrieval Technique, ook nodig hebben; vooral als een tekst vergezeld moet gaan van een tutorial met directe instructie of uitwerking van het onderwerp in focus. Dit is om ervoor te zorgen dat de belangrijkste concepten niet worden verwaterd of vervangen door de instructie of antwoorden van de leraar.

Met andere woorden, sommigen hebben de ‘aanpak’ bekritiseerd als te simpel. Als leren complex is en cognitie complex is, hoe kan assessment zo eenvoudig zijn? Ons antwoord is dat het vermogen om concepten uit het geheugen op te halen het eenvoudige bijproduct is van weliswaar complexe processen. Het verwerken van informatie door een computer kan ingewikkeld

zijn, maar het ophalen van informatie uit een resulterende database is (meestal) eenvoudig. Anderen wijzen op het feit dat onderwijs ideaal de studenten hogere cognitieve vaardigheden biedt, zoals het vermogen om passende problemen op te lossen, niet alleen kennis. Hoewel we het eens zijn met dit standpunt, willen we benadrukken dat probleem oplossing, net als andere cognitieve processen van hogere orde, zoals besluitvorming en beoordeling, volledig op kennis is gebaseerd; zonder de juiste kennis kan geen probleem worden opgelost, zoals studies in menselijke expertise op hoog niveau keer op keer hebben aangetoond (K.A. Ericsson, Charness, Feltovich, & Hoffman, 2006). (Dit wil natuurlijk niet zeggen dat het hebben van de juiste kennis een voldoende voorwaarde is om expertise op te doen, en niet alleen te verwerven, maar studenten ook hun kennis moeten toepassen in probleem oplossende situaties, maar dit vereist praktijk gerichte training en gerelateerde beoordelingen). Ten derde zou men kunnen betogen dat niet alle concepten even belangrijk zijn en dat de Concept Retrieval Technique deze stand van zaken niet adequaat weergeeft. Dit bezwaar kan gemakkelijk worden tegengegaan door te suggereren dat leraren moeten aangeven welke concepten in hun doelwoordenlijst een hoger gewicht moeten krijgen en eindscores moeten berekenen, rekening houdend met of de aanwezigheid of afwezigheid van deze concepten. Op basis van onze ervaring voorspellen we echter dat het wege van antwoorden geen groot verschil zal maken in de totale scores van studenten.



# References

## REFERENCES

**A**

- Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22(3), 261–295. [https://doi.org/10.1016/S0022-5371\(83\)90201-3](https://doi.org/10.1016/S0022-5371(83)90201-3)
- Atkinson, R. C., & Shiffrin, R. M. (1968). Human Memory: A Proposed System and its Control Processes. In *Psychology of Learning and Motivation - Advances in Research and Theory* (Vol. 2, pp. 89–195). [https://doi.org/10.1016/S0079-7421\(08\)60422-3](https://doi.org/10.1016/S0079-7421(08)60422-3)
- Attali, Y. (2015). Reliability-Based Feature Weighting for Automated Essay Scoring. *Applied Psychological Measurement*, 39(4), 303–313. <https://doi.org/10.1177/0146621614561630>
- Attali, Y., & Burstein, J. (2006). Automated Essay Scoring With e-rater V.2. *Journal of Technology, Learning, and Assessment*, 4(3). Retrieved from <https://www.learntechlib.org/p/103244/>
- Attali, Y., Lewis, W., & Steier, M. (2013). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing*, 30(1), 125–141. <https://doi.org/10.1177/0265532212452396>
- Attali, Y., & Powers, D. (2009). Validity of Scores for a Developmental Writing Scale Based on Automated Scoring. *Educational and Psychological Measurement*, 69(6), 978–993. <https://doi.org/10.1177/0013164409332217>
- Azevedo, R., & Bernard, R. M. (1995). A Meta-Analysis of the Effects of Feedback in Computer-Based Instruction. *Journal of Educational Computing Research*, 13(2), 111–127. <https://doi.org/10.2190/9LMD-3U28-3A0G-FTQT>

**B**

- Baddeley, A. (1992). Working Memory: The Interface between Memory and Cognition. *Journal of Cognitive Neuroscience*, 4(3), 281–288. <https://doi.org/10.1162/jocn.1992.4.3.281>
- Bejar, I. I. (2011). A validity based approach to quality control and assurance of automated scoring. *Assessment in Education: Principles, Policy & Practice*, 18(3), 319–341. <https://doi.org/10.1080/0969594x.2011.555329>
- Bejar, I. I., Williamson, D. M., & Mislevy, R. J. (2006). *Automated scoring of complex tasks in computer-based testing: An introduction*. Mahwah, NJ: Lawrence Erlbaum.
- Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, 85(5), 536–553. <https://doi.org/10.1002/sce.1022>
- Ben-Simon, A., & Bennett, R. E. (2007). Toward More Substantively Meaningful Automated Essay Scoring. *The Journal of Technology, Learning, and Assessment*, 6(1), 1–47.

Retrieved from <https://www.learntechlib.org/p/103252/>

- Bennett, R. E., & Bejar, I. I. (1998). Validity and Automated Scoring: It's Not Only the Scoring. *Educational Measurement: Issues and Practice*, 17(4), 9–17. <https://doi.org/10.1111/j.1745-3992.1998.tb00631.x>
- Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, 15(11), 527–536. <https://doi.org/10.1016/j.tics.2011.10.001>
- Birenbaum, M., & Feldman, R. A. (1998). Relationships between learning patterns and attitudes towards two assessment formats. *Educational Research*, 40(1), 90–98. <https://doi.org/10.1080/0013188980400109>
- Bordage, G. (1994). Elaborated knowledge. *Academic Medicine*, 69(11), 883–5. <https://doi.org/10.1097/00001888-199411000-00004>
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, 111(4), 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Brachman, R. J. (1977). What's in a concept: structural foundations for semantic networks. *International Journal of Man-Machine Studies*, 9(2), 127–152. [https://doi.org/10.1016/S0020-7373\(77\)80017-5](https://doi.org/10.1016/S0020-7373(77)80017-5)
- Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country. *Applied Measurement in Education*, 25(1), 27–40. <https://doi.org/10.1080/08957347.2012.635502>
- Brown, G. A., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education*. London: Routledge.
- Buckner, R. L., Wheeler, M. E., & Sheridan, M. A. (2001). Encoding Processes during Retrieval Tasks. *Journal of Cognitive Neuroscience*, 13(3), 406–415. <https://doi.org/10.1162/08989290151137430>
- Butcher, P. G., & Jordan, S. E. (2010). A comparison of human and computer marking of short free-text student responses. *Computers & Education*, 55(2), 489–499. <https://doi.org/10.1016/j.compedu.2010.02.012>

## C

- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/13634291>
- Campoy, G., Castellà, J., Provencio, V., Hitch, G. J., & Baddeley, A. (2015). Automatic semantic encoding in verbal short-term memory: Evidence from the concreteness effect.



- Quarterly Journal of Experimental Psychology*, 68(4), 759–778.  
<https://doi.org/10.1080/17470218.2014.966248>
- Carmines, E., & Zeller, R. (1979). *Reliability and Validity Assessment*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc.  
<https://doi.org/10.4135/9781412985642>
- Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, 20(6), 633–642. <https://doi.org/10.3758/BF03202713>
- Champagne, A. B., Klopfer, L. E., Desena, A. T., & Squires, D. A. (1981). Structural representations of students' knowledge before and after science instruction. *Journal of Research in Science Teaching*, 18(2), 97–111. <https://doi.org/10.1002/tea.3660180202>
- Chang, T. M. (1986). Semantic memory: Facts and models. *Psychological Bulletin*, 99(2), 199–220. <https://doi.org/10.1037/0033-2909.99.2.199>
- Charlin, B., Tardif, J., & Boshuizen, H. P. A. (2000). Scripts and Medical Diagnostic Knowledge. *Academic Medicine*, 75(2), 182–190. <https://doi.org/10.1097/00001888-200002000-00020>
- Chen, Z., & Cowan, N. (2005). Chunk Limits and Length Limits in Immediate Recall: A Reconciliation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(6), 1235–1249. <https://doi.org/10.1037/0278-7393.31.6.1235>
- Clauser, B. E., Kane, M. T., & Swanson, D. B. (2002). Validity Issues for Performance-Based Tests Scored With Computer-Automated Scoring Systems. *Applied Measurement in Education*, 15(4), 413–432. [https://doi.org/10.1207/S15324818AME1504\\_05](https://doi.org/10.1207/S15324818AME1504_05)
- Collins, A. M., & Loftus, E. F. (1975). A Spreading-Activation Theory of Semantic Processing. In *Readings in Cognitive Science* (Vol. 82, pp. 407–428). Elsevier.  
<https://doi.org/10.1016/B978-1-4832-1446-7.50015-7>
- Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2), 240–247. [https://doi.org/10.1016/S0022-5371\(69\)80069-1](https://doi.org/10.1016/S0022-5371(69)80069-1)
- Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, 18(1), 100–108.  
<https://doi.org/10.1016/j.asw.2012.11.001>
- Connors, R. J., & Lunsford, A. A. (1988). Frequency of Formal Errors in Current College Writing, or Ma and Pa Kettle Do Research. *College Composition and Communication*, 39(4), 395. <https://doi.org/10.2307/357695>
- Conrad, C. (1972). Cognitive economy in semantic memory. *Journal of Experimental*

*Psychology*, 92(2), 149–154. <https://doi.org/10.1037/h0032072>

Cowan, N. (2010). The Magical Mystery Four. *Current Directions in Psychological Science*, 19(1), 51–57. <https://doi.org/10.1177/0963721409359277>

Cushing Weigle, S. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing*, 27(3), 335–353. <https://doi.org/10.1177/0265532210364406>

## D

Daley, B. J., & Torre, D. M. (2010). Concept maps in medical education: an analytical literature review. *Medical Education*, 44(5), 440–448. <https://doi.org/10.1111/j.1365-2923.2010.03628.x>

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1), 7–24. <https://doi.org/10.1016/j.asw.2012.10.002>

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal Of Technology Learning And Assessment*, 5(1), 2006–12. Retrieved from <http://www.jtla.org>

## E

Eddleman, S. (2007). *CPO Focus on Life Science*. New Hampshire: CPO Science.

Edmondson, K. M. (2005). Assessing science understanding through concept maps. In *Assessing Science Understanding* (pp. 15–40). Elsevier. <https://doi.org/10.1016/B978-012498365-6/50004-4>

Eichenbaum, H. (2017). Memory: Organization and Control. *Annual Review of Psychology*, 68(1), 19–45. <https://doi.org/10.1146/annurev-psych-010416-044131>

Eppler, M. J. (2006). A Comparison between Concept Maps, Mind Maps, Conceptual Diagrams, and Visual Metaphors as Complementary Tools for Knowledge Construction and Sharing. *Information Visualization*, 5(3), 202–210. <https://doi.org/10.1057/palgrave.ivs.9500131>

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (2006). *The Cambridge Handbook of Expertise and Expert Performance*. (K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman, Eds.). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511816796>

## F

Flor, M., & Fugati, Y. (2012). On using context for automatic correction of non-word misspellings in student essays. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 105–115). Association for Computational Linguistic. Retrieved from <http://aclweb.org/anthology/W/W12/W12-2012.pdf>

Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing*, 20(4), 384–408. <https://doi.org/10.1191/0265532203lt265oa>

## G

Gay, L. R. (1980). The comparative effects of multiple-choice versus short-answer tests on retention. *Journal of Educational Measurement*, 17(1), 45–50.  
<https://doi.org/10.1111/j.1745-3984.1980.tb00813.x>

Glaser, R., & Bassok, M. (1989). Learning theory and the study of instruction. *Annual Review of Psychology*, 40, 631–666.

Glass, A. L., & Sinha, N. (2013). Multiple-Choice Questioning Is an Efficient Instructional Methodology That May Be Widely Implemented in Academic Courses to Improve Exam Performance. *Current Directions in Psychological Science*, 22(6), 471–477.  
<https://doi.org/10.1177/0963721413495870>

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory and Cognition*, 40(4), 505–513.  
<https://doi.org/10.3758/s13421-011-0174-0>

## H

Hebb, D. O. (1949). *The organization of behavior : a neuropsychological theory*. New York : Wiley.

Ho, V., Kumar, R. K., & Velan, G. (2014). Online testable concept maps: benefits for learning about the pathogenesis of disease. *Medical Education*, 48(7), 687–697.  
<https://doi.org/10.1111/medu.12422>

## J

Jayashankar, S., & Sridaran, R. (2017). Superlative model using word cloud for short answers evaluation in eLearning. *Education and Information Technologies*, 22(5), 2383–2402.  
<https://doi.org/10.1007/s10639-016-9547-0>

Joanisse, M. F., & McClelland, J. L. (2015). Connectionist perspectives on language learning, representation and processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(3), 235–247. <https://doi.org/10.1002/wcs.1340>

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). *Structural Knowledge: Techniques for Representing, Conveying, and Acquiring Structural Knowledge*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Jones, M. N., Willits, J., & Dennis, S. (2015). *Models of Semantic Memory*. (J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels, Eds.), *Oxford Handbook of Mathematical and Computational Psychology* (Vol. 1). Oxford University Press.

<https://doi.org/10.1093/oxfordhb/9780199957996.013.11>

- Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130–144.  
<https://doi.org/10.1016/j.edurev.2007.05.002>

## K

- Kersting, N. B., Sherin, B. L., & Stigler, J. W. (2014). Automated Scoring of Teachers' Open-Ended Responses to Video Prompts. *Educational and Psychological Measurement*, 74(6), 950–974. <https://doi.org/10.1177/0013164414521634>
- Kiefer, M., & Pulvermüller, F. (2012). Conceptual representations in mind and brain: Theoretical developments, current evidence and future directions. *Cortex*, 48(7), 805–825.  
<https://doi.org/10.1016/j.cortex.2011.04.006>
- Koponen, I. T., & Pehkonen, M. (2010). Coherent Knowledge Structures of Physics Represented as Concept Networks in Teacher Education. *Science & Education*, 19(3), 259–282.  
<https://doi.org/10.1007/s11191-009-9200-z>
- Krause, U.-M., Stark, R., & Mandl, H. (2009). The effects of cooperative learning and feedback on e-learning in statistics. *Learning and Instruction*, 19(2), 158–170.  
<https://doi.org/10.1016/j.learninstruc.2008.03.003>

## L

- Landis, J. R., & Koch, G. G. (1977). An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics*, 33(2), 363.  
<https://doi.org/10.2307/2529786>
- Leiva, F. M., Ríos, F. J. M., & Martínez, T. L. (2006). Assessment of Interjudge Reliability in the Open-Ended Questions Coding Process. *Quality & Quantity*, 40(4), 519–537.  
<https://doi.org/10.1007/s11135-005-1093-6>
- Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, 53(2), 215–233.  
<https://doi.org/10.1002/tea.21299>

## M

- Martin, A. (2007). The Representation of Object Concepts in the Brain. *Annual Review of Psychology*, 58(1), 25–45. <https://doi.org/10.1146/annurev.psych.57.102904.190143>
- Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: structure and processes. *Current Opinion in Neurobiology*, 11(2), 194–201. [https://doi.org/10.1016/S0959-4388\(00\)00196-3](https://doi.org/10.1016/S0959-4388(00)00196-3)

- McClure, J. R., Sonak, B., & Suen, H. K. (1999). Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching*, 36(4), 475–492. [https://doi.org/10.1002/\(SICI\)1098-2736\(199904\)36:4<475::AID-TEA5>3.0.CO;2-O](https://doi.org/10.1002/(SICI)1098-2736(199904)36:4<475::AID-TEA5>3.0.CO;2-O)
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Meyer, D. E. (1970). On the representation and retrieval of stored semantic information. *Cognitive Psychology*, 1(3), 242–299. [https://doi.org/10.1016/0010-0285\(70\)90017-4](https://doi.org/10.1016/0010-0285(70)90017-4)
- Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97. <https://doi.org/10.1037/h0043158>
- Muhlenbach, F., Lallich, S., & Zighed, D. A. (2004). No Title. *Journal of Intelligent Information Systems*, 22(1), 89–109. <https://doi.org/10.1023/A:1025832930864>
- N**
- Nicol, D. (2007). E-assessment by design: using multiple-choice tests to good effect. *Journal of Further and Higher Education*, 31(1), 53–64. <https://doi.org/10.1080/03098770601167922>
- Nnodim, J. O. (1992). Multiple-choice testing in anatomy. *Medical Education*, 26(4), 301–309. <https://doi.org/10.1111/j.1365-2923.1992.tb00173.x>
- Noorbehbahani, F., & Kardan, A. A. (2011). The automatic assessment of free text answers using a modified BLEU algorithm. *Computers & Education*, 56(2), 337–345. <https://doi.org/10.1016/j.compedu.2010.07.013>
- Novak, J. D. (1990). Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching*, 27(10), 937–949. <https://doi.org/10.1002/tea.3660271003>
- Novak, J. D., & Gowin, B. D. (1984). *Learning how to learn*. New York: Cambridge University Press.
- P**
- Patel, V. L., & Groen, G. J. (1986). Knowledge Based Solution Strategies in Medical Reasoning. *Cognitive Science*, 10(1), 91–116. [https://doi.org/10.1207/s15516709cog1001\\_4](https://doi.org/10.1207/s15516709cog1001_4)
- Pyc, M. A., & Rawson, K. A. (2010). Why Testing Improves Memory: Mediator Effectiveness Hypothesis. *Science*, 330(6002), 335–335. <https://doi.org/10.1126/science.1191465>
- Q**
- Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, 12(5), 410–430. <https://doi.org/10.1002/bs.3830120511>

## R

- Raaijmakers, J. G., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review*, 88(2), 93–134. <https://doi.org/10.1037/0033-295X.88.2.93>
- Raaijmakers, J. G., & Shiffrin, R. M. (1992). Models for Recall and Recognition. *Annual Review of Psychology*, 43(1), 205–234. <https://doi.org/10.1146/annurev.ps.43.020192.001225>
- Ramsden, A., & Bate, A. (2008). *Using word clouds in teaching and learning*. University of Bath.
- Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, 35(4), 435–448. <https://doi.org/10.1080/02602930902862859>
- Redecker, C., & Johannessen, Ø. (2013). Changing Assessment - Towards a New Assessment Paradigm Using ICT. *European Journal of Education*, 48(1), 79–96. <https://doi.org/10.1111/ejed.12018>
- Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18–39. <https://doi.org/10.1016/j.asw.2010.01.003>
- Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, 12(1), 1–20. [https://doi.org/10.1016/S0022-5371\(73\)80056-8](https://doi.org/10.1016/S0022-5371(73)80056-8)
- Roberts, V., & Joiner, R. (2007). Investigating the efficacy of concept mapping with pupils with autistic spectrum disorder. *British Journal of Special Education*, 34(3), 127–135. <https://doi.org/10.1111/j.1467-8578.2007.00468.x>
- Rodriguez, M. C. (2005). Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research. *Educational Measurement: Issues and Practice*, 24(2), 3–13. <https://doi.org/10.1111/j.1745-3992.2005.00006.x>
- Roediger, H. L., & Guynn, M. J. (1996). Retrieval Processes. In *Memory* (pp. 197–236). Elsevier. <https://doi.org/10.1016/B978-012102570-0/50009-4>
- Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition : A parallel distributed processing approach*. Cambridge, MA, US: MIT Press.
- Rogers, T. T., & McClelland, J. L. (2008). Précis of Semantic Cognition: A Parallel Distributed Processing Approach. *Behavioral and Brain Sciences*, 31(06), 689. <https://doi.org/10.1017/S0140525X0800589X>
- Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of

- categories. *Cognitive Psychology*, 7(4), 573–605. [https://doi.org/10.1016/0010-0285\(75\)90024-9](https://doi.org/10.1016/0010-0285(75)90024-9)
- Rotgans, J. I., & Schmidt, H. G. (2014). Situational interest and learning: Thirst for knowledge. *Learning and Instruction*, 32, 37–50. <https://doi.org/10.1016/j.learninstruc.2014.01.002>
- Ruiz-Primo, M. A., Schultz, S. E., Li, M., & Shavelson, R. J. (2001). Comparison of the reliability and validity of scores from two concept-mapping techniques. *Journal of Research in Science Teaching*, 38(2), 260–278. [https://doi.org/10.1002/1098-2736\(200102\)38:2<260::AID-TEA1005>3.0.CO;2-F](https://doi.org/10.1002/1098-2736(200102)38:2<260::AID-TEA1005>3.0.CO;2-F)
- Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching*, 33(6), 569–600. [https://doi.org/10.1002/\(SICI\)1098-2736\(199608\)33:6<569::AID-TEA1>3.0.CO;2-M](https://doi.org/10.1002/(SICI)1098-2736(199608)33:6<569::AID-TEA1>3.0.CO;2-M)
- Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. *Attention and Performance. XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*, 3–30.
- S**
- Schmidt, H. G., Rotgans, J. I., & Yew, E. H. (2011). The process of problem-based learning: what works and why. *Medical Education*, 45(8), 792–806. <https://doi.org/10.1111/j.1365-2923.2011.04035.x>
- Schmidt, H. G., Van Der Molen, H. T., Te Winkel, W. W. R., & Wijnen, W. H. F. W. (2009). Constructivist, Problem-Based Learning Does Work: A Meta-Analysis of Curricular Comparisons Involving a Single Medical School. *Educational Psychologist*, 44(4), 227–249. <https://doi.org/10.1080/00461520903213592>
- Schunk, D. H. (2008). Metacognition, Self-Regulation, and Self-Regulated Learning: Research Recommendations. *Educational Psychology Review*, 20(4), 463–467. <https://doi.org/10.1007/s10648-008-9086-3>
- Schuwirth, L. W. T., & Van der Vleuten, C. P. M. (2011). Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher*, 33(6), 478–485. <https://doi.org/10.3109/0142159X.2011.565828>
- Shavelson, R. J. (1972). Some aspects of the correspondence between content structure and cognitive structure in physics instruction. *Journal of Educational Psychology*, 63(3), 225–234. <https://doi.org/10.1037/h0032652>
- Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20, 53–76. <https://doi.org/10.1016/j.asw.2013.04.001>

- Shermis, M. D., Page, E. B., & Keith, T. Z. (2002). Types of Grading Engines. *Educational and Psychological Measurement*, 62(1), 5–18.
- Shiffrin, R. M., & Nosofsky, R. M. (1994). Seven plus or minus two: A commentary on capacity limitations. *Psychological Review*, 101(2), 357–361.
- Shuell, T. J. (1986). Cognitive Conceptions of Learning. *Review of Educational Research*, 56(4), 411–436. <https://doi.org/10.3102/00346543056004411>
- Siddiqi, R., Harrison, C. J., & Siddiqi, R. (2010). Improving Teaching and Learning through Automated Short-Answer Marking. *IEEE Transactions on Learning Technologies*, 3(3), 237–249. <https://doi.org/10.1109/TLT.2010.4>
- Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 81(3), 214–241. <https://doi.org/10.1037/h0036351>
- Stellmack, M. A., Konheim-Kalkstein, Y. L., Manor, J. E., Massey, A. R., & Schmitz, J. A. P. (2009). An Assessment of Reliability and Validity of a Rubric for Grading APA-Style Introductions. *Teaching of Psychology*, 36(2), 102–107. <https://doi.org/10.1080/00986280902739776>
- Sternberg, R. J. (1984). A theory of knowledge acquisition in the development of verbal concepts\*1. *Developmental Review*, 4(2), 113–138. [https://doi.org/10.1016/0273-2297\(84\)90001-7](https://doi.org/10.1016/0273-2297(84)90001-7)
- Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition*, 38(2), 244–253. <https://doi.org/10.3758/MC.38.2.244>
- T**
- Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education*, 9(1), 40. <https://doi.org/10.1186/1472-6920-9-40>
- Tennyson, R. D., & Rasch, M. (1988). Linking cognitive learning theory to instructional prescriptions. *Instructional Science*, 17(4), 369–385. <https://doi.org/10.1007/BF00056222>
- Thompson-Schill, S. L. (2003). Neuroimaging studies of semantic memory: inferring “how” from “where.” *Neuropsychologia*, 41(3), 280–292. [https://doi.org/10.1016/S0028-3932\(02\)00161-6](https://doi.org/10.1016/S0028-3932(02)00161-6)
- Tulving, E. (1986). Episodic and semantic memory: Where should we go from here? *Behavioral and Brain Sciences*, 9(03), 573. <https://doi.org/10.1017/S0140525X00047257>



Tulving, E. (1993). What Is Episodic Memory? *Current Directions in Psychological Science*, 2(3), 67–70. <https://doi.org/10.1111/1467-8721.ep10770899>

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4), 327–352. <https://doi.org/10.1037/0033-295X.84.4.327>

## U

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, 114(1), 104–132. <https://doi.org/10.1037/0033-295X.114.1.104>

## V

Van den Broek, G., Takashima, A., Wiklund-Hörnqvist, C., Karlsson Wirebring, L., Segers, E., Verhoeven, L., & Nyberg, L. (2016). Neurocognitive mechanisms of the “testing effect”: A review. *Trends in Neuroscience and Education*, 5(2), 52–66. <https://doi.org/10.1016/j.tine.2016.05.001>

Vinet, L., & Zhedanov, A. (2011). A ‘missing’ family of classical orthogonal polynomials. *Journal of Physics A: Mathematical and Theoretical*, 44(8), 085201. <https://doi.org/10.1088/1751-8113/44/8/085201>

## W

Wagner, A. D. (1998). Building Memories: Remembering and Forgetting of Verbal Experiences as Predicted by Brain Activity. *Science*, 281(5380), 1188–1191. <https://doi.org/10.1126/science.281.5380.1188>

Wass, V., Van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *The Lancet*, 357(9260), 945–949. [https://doi.org/10.1016/S0140-6736\(00\)04221-5](https://doi.org/10.1016/S0140-6736(00)04221-5)

West, D. C., Pomeroy, J. R., Park, J. K., Gerstenberger, E. A., & Sandoval, J. (2000). Critical Thinking in Graduate Medical Education. *JAMA*, 284(9), 1105. <https://doi.org/10.1001/jama.284.9.1105>

Wolfe, P. (2001). *Brain matters: translating research into classroom practice*. Alexandria, Va: Association for Supervision and Curriculum Development. Retrieved from <http://ebookcentral.proquest.com/lib/columbia/detail.action?docID=624063>

## Y

Yew, E. H. J., Chng, E., & Schmidt, H. G. (2011). Is learning in problem-based learning cumulative? *Advances in Health Sciences Education*, 16(4), 449–464. <https://doi.org/10.1007/s10459-010-9267-y>

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education*, 45(4), 477–501.  
<https://doi.org/10.1023/A:1023967026413>

## **Z**

Zhang, M. (2013). Contrasting automated and human scoring of essays. *ETS R & D Connections*, (21), 11 p. Retrieved from  
[https://www.ets.org/Media/Research/pdf/RD\\_Connections\\_21.pdf](https://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf)





# Curriculum Vitae

## CURRICULUM VITAE

Gavin Hays has been at the forefront of the pedagogical change in Australia secondary schools, specifically in the areas of PBL, flipped learning and STEM. He completed a Bachelor of Arts and Bachelor of Teaching (Secondary) from Australia Catholic University (Sydney) in 2002. Starting his teaching career at Parramatta Marist High School, where he undertook a number of leadership roles including Leader of Pedagogy and Assistant Principal. Fortunately, during this period of employment he was afforded a number of international professional development opportunities within the United States, Singapore and Finland. During this time, he achieved accreditation as a PBL trainer within the New Tech Network and recognition as a Problem Crafter within the Republic Polytechnic in Singapore.

In 2008, he earned a master's degree in Educational Leadership from Australia Catholic University (Sydney). In 2015, he started his PhD study with Professor Henk Schmidt and Associate Professor Dr Jerome Rotgans, exploring the implementation of the Concept Retrieval Technique in primary and secondary active learning environments. With a keen interest in the use of automated assessment, he has undertaken the development of a pilot software program that automates the scoring of the technique. During his PhD study, he has presented his work at several international conferences including the PBL Congress, Zurich, Switzerland and American Educational Research Association (AERA) Annual Conference, San Antonio, USA.

He now works with the Catholic Education Catholic Diocese of Parramatta as a Learning Leader to support schools as they navigate change in implementing contemporary learning pedagogies. Finally, he is currently serving on the executive of the Australian Curriculum Studies Association (ACSA) who advocate for curriculum development and professional learning across all educational sectors.



# **Author Publications**

## AUTHOR PUBLICATIONS

- Hendry, A., Hays, G., Challinor, K. & Lynch, D. (2016). Enhancing student learning through Project Based Learning (PBL) in a secondary school integrative STEM course. In: 27th Annual Conference of the Australasian Association for Engineering Education: AAEE 2016. Lismore, NSW: Southern Cross University, 2016: 337-348. ISBN: 9780994152039.
- Hendry, A., Hays, G., Challinor, K., & Lynch, D. (2017). Undertaking Educational Research Following the Introduction, Implementation, Evolution, and Hybridization of Constructivist Instructional Models in an Australian PBL High School. *Interdisciplinary Journal of Problem-Based Learning*, 11(2). <https://doi.org/10.771/1541-5015.1688>.
- Hendry, A. & Hays, G. (2018). Changing Education in Action: Lighting the Collective Efficacy Flame. In J. Andrews, D. Netolicky & C. Paterson (Eds), *Flip the system in Australia: What matters in education*. London, United Kingdom: Taylor & Francis.

