



MONASH University

Development and Validation of an Instrument to Measure Undergraduate Chemistry Students' Critical Thinking Skills

Stephen Michael Danczak
Bachelor of Science (Honours)

A thesis submitted for the degree of Doctor of Philosophy at
Monash University in 2018
School of Chemistry, Faculty of Science

Copyright notice

© The author 2018.

I certify that I have made all reasonable efforts to secure copyright permissions for third-party content included in this thesis and have not knowingly added copyright content to my work without the owner's permission.

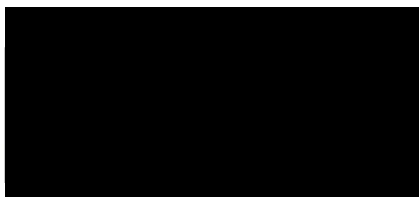
Abstract

The importance of developing student critical thinking at university can be seen through its inclusion as a graduate attribute for universities and from research highlighting the value employers, educators and students place on demonstrating critical thinking skills. Commercially available critical thinking assessments are generic in context, however assessments using a context relevant to the student are believed to more accurately reflect the students' critical thinking skills. This thesis describes the development and evaluation of Danczak-Overton-Thompson chemistry critical thinking test (DOT-CCTT), set in a chemistry context, and designed to administer to undergraduate chemistry students at any level of study. Prior to developing the DOT-CCTT, a qualitative study revealed students' definitions of critical thinking centred on 'problem solving' and 'analysis', and they believed that they developed critical thinking skills in inquiry-style laboratory environments. Development and evaluation occurred over three versions of the DOT-CCTT through a variety of reliability and validity testing. The studies suggest that the final version of the DOT-CCTT has good internal reliability, strong test-retest reliability, moderate convergent validity relative to a commercially available test, and is independent of previous academic achievement and university of study. Criterion validity testing revealed that third year students performed statistically significantly better on the DOT-CCTT relative to first year students, and postgraduates and academics performed statistically significantly better than third year students. Qualitative analysis also provided evidence of peer learning which suggests the DOT-CCTT may be a valuable teaching tool. The statistical and qualitative analysis indicates that the DOT-CCTT is a suitable instrument for the chemistry education community to use to measure the development of undergraduate chemistry students' critical thinking skills. Alternatively the DOT-CCTT may be useful as a discussion tool to assist in the development of undergraduate chemistry students' critical thinking skills.

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Signature:

A solid black rectangular box used to redact the signature.

Print Name: Stephen Michael Danczak

Date: 12/03/2018

Publications during enrolment

Danczak, S. M., Thompson, C. D., & Overton, T. L. (2017). What does the term critical thinking mean to you? A qualitative analysis of chemistry undergraduate, teaching staff and employers' views of critical thinking. *Chemistry Education Research and Practice*, **18**(3), 420-434.

Thesis including published works declaration

This thesis includes one original paper published in peer reviewed journals and zero submitted publications. The core theme of the thesis is the development and validation of an instrument to measure undergraduate chemistry students' critical thinking skills. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the School of Chemistry under the supervision of Dr Christopher David Thompson.

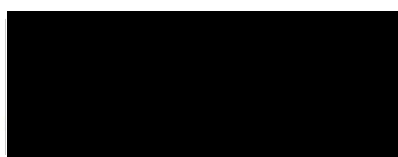
(The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.)

In the case of chapter 2 my contribution to the work involved the following:

Thesis Chapter	Publication Title	Status (published, in press, accepted or returned for revision, submitted)	Nature and % of student contribution	Co-author name(s) Nature and % of Co-author's contribution*	Co-author(s), Monash student Y/N*
2	'What does the term Critical Thinking mean to you?' A qualitative analysis of chemistry undergraduate, teaching staff and employers' views of critical thinking.	Accepted	95%. Concept and collecting data and writing first draft	Christopher D Thompson, input into manuscript 33% Tina L Overton, input into manuscript 33%	No No

I have not renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

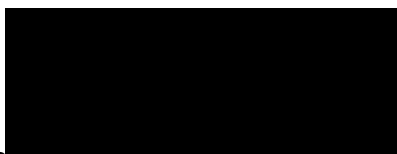
Student signature:



Date: 12/03/2018

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor signature



Date: 12/03/2018

Acknowledgements

First and foremost I want to thank my supervisors Dr Chris Thompson and Prof Tina Overton for their unyielding support, advice, encouragement and friendship. Dealing with mental issues has been a core part of my growth through my PhD journey and you have shown me infinite patience and words cannot begin to explain how grateful I am to you both. Chris, thank you so much for sharing this world of chemistry education research with me, and Tina, thank you for all the tireless years you put into helping forge the discipline of chemistry education research into what it is today. I hope I can continue to make you both proud.

I'd like to briefly thank the past and present (that includes Aishling) members of the Monash Chemistry Education and Research Group and Team Goldfish. Thank you for tolerating my sense of humour which is largely centred on pop-culture, puns and profanity. You have all been my daily support and friends throughout this process.

I'd like to thank all my research participants throughout my PhD for without your generous help I would have no data with which to analyse or write about. I would also like to thank the Australian government as this research was supported by an Australian Government Research Training Program (RTP) Scholarship.

To my parents, Olga and Michael, I don't know where to start with my gratitude. I have my entire life to thank you both for and I am grateful for your support in whatever I've done. Thank you for never giving up on me even when I had given up on myself. I've enjoyed our chats in recent years and teaching you both about the inner workings of universities and academia.

To my dearest Viv, thank you so much for your love, infectious drive and motivation. Without you, I'm not sure I would have made it this far. You've made such a huge effort to understand me and support me when everything seemed hopeless. I could write an entire thesis just on how much you mean to me, but suffice to say your love keeps lifting me high and higher. I love you so much.

And finally, I'd like to thank my psychologist Lewi, for his counsel and helping me through this journey to fight depression and anxiety to complete this thesis.

Table of Contents

Copyright notice	ii
Abstract.....	iii
Declaration	iv
Publications during enrolment	v
Thesis including published works declaration	vi
Acknowledgements	vii
Table of Contents.....	viii
List of Tables.....	xii
List of Figures.....	xiii
Glossary of Key Terms and Acronyms	xiv
Chapter 1 Introduction	1
1.1 How is Critical Thinking Defined?	1
1.2 How is Critical Thinking Taught?.....	4
1.2.1 Recent Approaches Used to Develop Critical Thinking.....	11
1.3 The Assessment of Critical Thinking	12
1.3.1 Approaches to Assessing Critical Thinking in Higher Education Chemistry	13
1.4 Commonly Used Critical Thinking Tests and Teaching Tools	15
1.4.1 Chemistry Specific Tests and Teaching Tools	22
1.5 Review of Critical Thinking Assessment Tools.....	24
1.6 Considerations When Evaluating Critical Thinking Skills.....	26
1.7 Research Question	27
Chapter 2 Methodology	31
2.1 Qualitative Research Theoretical Framework	31
2.2 Test Theory Framework.....	32
2.3 Reliability	34
2.3.1 Internal Reliability.....	36
2.3.2 Item Difficulty Analysis	36
2.3.3 Test-Retest Reliability	38

2.4	Validity.....	38
2.4.1	Convergent Validity	40
2.4.2	Criterion Validity	40
2.4.3	Discriminate Validity	41
2.4.4	Content Validity	41
2.5	Use of Non-Parametric Tests Versus Parametric Tests	42
2.5.1	Assumptions Regarding Type 1 and Type 2 Errors	42
2.6	Ethics.....	43
2.7	Methodology Overview	44
2.8	A Qualitative Analysis of Students', Teaching Staff and Employers' Views of Critical Thinking.....	45
2.9	Operational Definition of Critical Thinking	64
2.10	Development of the Danczak-Overton-Thompson Chemistry Critical Thinking Test (DOT-CCTT).....	66
2.11	DOT-CCTTv1 Internal Reliability and Content Validity	73
2.11.1	Undergraduate Student Participants.....	74
2.11.2	Academic and Teaching Designer Participants	74
2.11.3	Internal Reliability Method	74
2.11.4	Content Validity Method	76
2.12	DOT-CCTTv2 Test-Retest Reliability, Convergent Validity and Content Validity	76
2.12.1	Undergraduate Student Participants.....	78
2.12.2	Structure of the Two Day Study.....	79
2.12.3	Test-Retest Reliability and Convergent Validity Method	80
2.12.4	Content Validity Method	80
2.13	DOT-CCTTv3 Internal Reliability, Criterion Validity and Discriminate Validity	81
2.13.1	Undergraduate Student Participants.....	84
2.13.2	Post-Doctoral Researchers, Honours and Phd Student Participants	85
2.13.3	Academic Participants.....	86
2.13.4	Treatment of Data	86
2.13.5	Internal Reliability Method	88

2.13.6	Criterion Validity Method	88
2.13.7	Discriminate Validity Method	89
2.14	Chapter Summary.....	89
Chapter 3	Results and Discussion	91
3.1	Introduction.....	91
3.2	DOT-CCTTv1	92
3.2.1	Internal Reliability	92
3.2.2	Item Difficulty Analysis	93
3.2.3	Content Validity	97
3.3	DOT-CCTTv2	106
3.3.1	Demographic Data of the Participants	106
3.3.2	Test-Retest Reliability	107
3.3.3	Convergent Validity	109
3.3.4	Content Validity	110
3.3.5	DOT-CCTTv2: Reliability and Validity Summary.....	125
3.4	DOT-CCTTv3	125
3.4.1	Demographic Data of Participants	126
3.4.2	Internal Reliability	126
3.4.3	Item Difficulty Analysis	128
3.4.4	Criterion Validity	131
3.4.5	Discriminate Validity	139
3.5	DOT-CCTTv3: Reliability and Validity Summary	141
3.6	Implications for Practice and Further Work	141
Chapter 4	Conclusions	146
Chapter 5	References.....	151
Chapter 6	Appendices	165
Appendix A:	Monash University Human Research Ethics Committee Certificate for Qualitative Research Investing Student's, Teaching Staff and Employer's Understanding of Critical Thinking	165

Appendix B: Monash University Human Research Ethics Committee Certificate for Research of the Danczak-Overton-Thompson Chemistry Critical Thinking Test (DOT-CCTT)	166
Appendix C: Student Questionnaire Regarding Understanding of Critical Thinking	167
Appendix D: Student Questionnaire Regarding Understanding of Critical Thinking	169
Appendix E: Danczak-Overton-Thompson Chemistry Critical Thinking Test Version 1 (DOT-CCTTv1) with Answers	171
Appendix F: DOT-CCTTv1 Statistics Summary	182
Appendix G: DOT-CCTTv1 Sample Academic Transcript and Coding	185
Appendix H: Danczak-Overton-Thompson Chemistry Critical Thinking Test Version 2 (DOT-CCTTv2) with Answers	186
Appendix I: DOT-CCTTv2 Statistics Summary	197
Appendix J: CCTTv2 Sample Student transcript and coding	200
Appendix K: Danczak-Overton-Thompson Chemistry Critical Thinking Test Version 3 (DOT-CCTTv3) with Answers.	203
Appendix L: DOT-CCTTv3 Statistics Summary	218

List of Tables

Table 1. 1	Categorization of critical thinking skills as defined by the philosophy, cognitive psychology and education research disciplines.....	4
Table 1. 2	Summary of commonly used commercially available critical thinking tests.....	17
Table 2. 1	Summary of participants' level of tertiary education/employment and the resulting 'Education group' used for statistical analysis.....	87
Table 3. 1	Summary of internal reliability data for the DOT-CCTTv1.....	93
Table 3. 2	Themes identified in the qualitative analysis of the academic focus groups for DOT-CCTTv1.....	98
Table 3. 3	Themes identified in the qualitative analysis of the student focus groups for the DOT-CCTTv2 and the WGCTA-S	113
Table 3. 4	Demographic data of participants who attempted the DOT-CCTTv3.....	126
Table 3. 5	Summary of internal reliability data for the DOT-CCTTv3.....	127
Table 3. 6	Mann-Whitney <i>U</i> tests comparing the median scores obtained on the DOT-CCTTv3 of each education group	131
Table 3. 7	Percentage of questions answered correctly on the DOT-CCTTv3 according to education group	133
Table 3. 8	Percentage of participants who answered questions 1 to 7 (Section 1: Making Assumptions) of the DOT-CCTTv3 correctly according to education group....	134
Table 3. 9	Percentage of participants who answered questions 8 to 13 (Section 2: Developing Hypotheses) of the DOT-CCTTv3 correctly according to education group	134
Table 3. 10	Percentage of participants who answered questions 14 to 18 (Section 3: Testing Hypotheses) of the DOT-CCTTv3 correctly according to education group	135
Table 3. 11	Percentage of participants who answered questions 19 to 23 (Section 4: Drawing Conclusions) of the DOT-CCTTv3 correctly according to education group	135
Table 3. 12	Percentage of participants who answered questions 24 to 30 (Section 5: Analysing Arguments) of the DOT-CCTTv3 correctly according to education group	135

List of Figures

Figure 1. 1	Timeline highlighting the evolution of the role of discipline specific knowledge in critical thinking. Distances on the horizontal axis indicates proximity from generalist or specifist views.	11
Figure 2. 1	Flow chart of methodology, consisting of writing and evaluating the reliability and validity of iterations of the DOT-CCTT	44
Figure 2. 2	Diagrammatic representation of how the operational definition of critical thinking within this study. The operational definition is situated between the literature definitions of critical thinking and student definition determined by Danczak <i>et al.</i> (2017).....	64
Figure 3. 1	Percentage of students who answered questions correct on DOT-CCTTv1...94	
Figure 3. 2	Median score of participants who answered each DOT-CCTTv1 correctly versus incorrectly.....	95
Figure 3. 3	Effect sizes (r) of the difference between median scores of participants who answered each DOT-CCTTv1 question correct or incorrect.....	96
Figure 3. 4	Percentage of students who answered questions correct on DOT-CCTTv3.129	
Figure 3. 5	Effect sizes (r) of the difference between median scores of participants who answered each DOT-CCTTv3 question correct or incorrect.....	130

Glossary of Key Terms

Corrected item total correlations: The correlation of each question relative to the score of a sub-scale on a psychometric test. A value $> .3$ is desirable

Content validity: Validity measure of a psychometric test concerned with the appropriateness of questions to correctly measure the psychometric construct of interest

Convergent validity: Validity measure of a psychometric test relative to a test of proven reliability and validity which also measures the same psychometric construct of interest

Criterion validity: Validity measure of a psychometric test's ability to predict the behaviour of the test taker relating to the psychometric construct of interest

Cronbach's α : Statistical value which is determined from averaging the correlations of all questions/sub-scales within a test against all other questions/sub-scales within the same test. Typically, a value of $\alpha > .7$ suggests the questions/sub-scales are measuring the same psychometric construct of interest

Discriminate validity: Validity measure of a psychometric test's dependence of discriminate variables such as age or sex

Effect size (r): The strength of the correlation between score and a discriminate variable such as sex. A value of $r > .5$ is considered a strong effect, whereas a value of $r < .1$ is considered a small effect

Generalist: Pedagogical view that critical thinking can be developed independent of discipline specific knowledge

Internal reliability: Reliability measure of a psychometric test used to determine whether questions/sub-scales of a test are measuring the same psychometric construct

Mann-Whitney U test: Non-parametric analysis used to determine any statistically significant difference between median scores of two groups

Non-parametric: Statistical data unevenly distributed either side of the mean in a non-bell curve fashion. Dictates appropriateness of statistical analyses

Spearman's ρ correlation (ρ): Non-parametric analysis used to determine the correlation between test scores and a discriminate variable. Correlation values can range from -1.00 to 1.00, indicating whether the variable and the score are inversely or directly correlated

Spearman-Brown Coefficient (ρ_{cc}): The internal reliability as measured by the correlations between test halves. A test is said to exhibit good internal reliability when $\rho_{cc} > .8$

Specifist: Pedagogical view that critical thinking cannot be developed independent of disciple specific knowledge

Statistical significance (p): The degree of overlap between variables. A value of $p > .05$ indicates less than a 5% overlap between variables due to chance

Test-retest reliability: Reliability measure of a psychometric test's ability to reproduce the same participant scores assuming no change has occurred in the participants relative to the construct of interest

Wilcoxon signed rank test: Non-parametric analysis used to determine whether a statistically significant difference between the test scores of the same participants taken at two time intervals

Glossary of Acronyms

ATAR:	Australian Tertiary Admission Rank
CCTDI:	California Critical Thinking Disposition Inventory
CCTST:	California Critical Thinking Skills Test
CCTT-Z:	Cornell Critical Thinking Test level Z
CITC:	Corrected item total correlations
DOT-CCTT:	Danczak-Overton-Thompson Chemistry Critical Thinking Test
EWCTET:	Ennis-Weir Critical Thinking Essay Test
HCTA:	Halpern Critical Thinking Appraisal
WGCTA:	Watson-Glaser Critical Thinking Appraisal
WGCTA-S:	Watson-Glaser Critical Thinking Appraisal Short Form

Chapter 1 Introduction

This thesis aimed to understand what is critical thinking, how is critical thinking taught, and how critical thinking is assessed. The context of this thesis was set within the teaching of chemistry at a tertiary level. This chapter serves as the introduction to the thesis as a whole. The chapter will commence with a discussion of how critical thinking is defined within the disciplines which contribute largely to academic literature of critical thinking, namely, philosophy, cognitive psychology and education research. Section 1.2 will introduce the teaching approaches used to develop critical thinking, firstly discussing the role of discipline specific knowledge with reference to the historical development of critical thinking, then highlighting pedagogical approaches. The discussion of critical thinking teaching approaches will then be followed by an introduction to the assessment of critical thinking in Section 1.3. Sections 1.4 and 1.5 will highlight popular testing approaches and review their strengths and weaknesses. Section 1.6 will build on the review of critical thinking assessment tests to discuss key considerations when evaluating critical thinking skills. Section 1.7 will then discuss the societal contexts which have driven the interest in developing the critical thinking skills of tertiary students and refine those arguments to construct the research question of this thesis.

1.1 How is Critical Thinking Defined?

Three disciplines dominate the discussion around the definition of critical thinking: philosophy, cognitive psychology and education research. The philosophy literature focuses on the generation of an argument or opinion (Facione, Sánchez, Facione and Gainen, 1995). The psychology literature considers problem solving and decision making as the essential outcomes of the critical thinking process (Halpern, 1996b). Whilst the education research literature is focused on observable behaviours (Barnett, 1997).

Among philosophers, one of the most commonly cited definitions of critical thinking is drawn from the Delphi Report (Facione, 1990). This report compiles a comprehensive dialogue regarding critical thinking between 47 academics from philosophy, education, social sciences and physical sciences. The consensus of these experts was that critical thinking is defined as

‘purposeful, self-regulatory judgement which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgement is based’ (Facione, 1990, p. 2). Despite being developed over 25 years ago this report is still relevant and the definition provided is still commonly used in recent literature (Abrami *et al.*, 2015; Desai, Berger and Higgs, 2016; Stephenson and Sadler-Mcknight, 2016).

The Delphi Report goes on to describe a series of cognitive skills and dispositions considered exemplary of an individual exhibiting good critical thinking. It suggests a good critical thinker is proficient in the skills of interpretation, analysis, evaluation, inference, explanation and self-regulation. It further suggests an individual may be in possession of such skills but without the necessary dispositions, the individual will not exhibit the behaviour of a good critical thinker: a ‘good critical thinker, is habitually disposed to engage in, and to encourage others to engage in, critical judgement’ (Facione, 1990, p. 12). These dispositions are broadly categorised into ‘approaches to life and living in general’ and ‘approaches to specific issues, questions or problems’. ‘Approaches to life and living in general’ is further categorised into 12 affective dispositions including inquisitiveness with regard to a wide range of issues, open-mindedness regarding divergent world views, self-confidence in one’s own ability reason, understanding the options of other people, and willingness to reconsider and revise views where honest reflection suggests the change is warranted. ‘Approaches to specific issues’ includes seven affective dispositions such as orderliness in working with complexity, diligence in seeking relevant information, persistence though difficulties are encountered, and precision to the degree permitted by the subject and circumstances (Facione, 1990).

Cognitive psychologists and education researchers use the term critical thinking to describe a set of cognitive skills, strategies or behaviours that increase the likelihood of a desired outcome (Halpern, 1996b; Tiruneh, Verburch and Elen, 2014). Psychologists typically investigate critical thinking experimentally and have developed a series of reasoning schemas with which to study and define critical thinking; conditional reasoning, statistical reasoning,

methodological reasoning and verbal reasoning (Lehman and Nisbett, 1990; Nisbett, Fong, Lehman and Cheng, 1987). Conditional reasoning draws from schemas where permission or an obligation for an event to occur is either satisfied or violated in the general form of premises such as 'if p, then q'. Statistical reasoning is the inferential reasoning applied when assessing the value of a sample relative to the population. Methodological reasoning is that which is employed when making judgements regarding sample and personal bias, errors in self-selected participants and the need for control groups. Finally verbal reasoning, is related to argument articulation and believed to improve with maturity.

Halpern (1993) expanded on these schemas to define critical thinking as the thinking required to solve problems, formulate inferences, calculate likelihoods and make decisions. Much like the philosophers, Halpern described a series of skills and dispositions of good critical thinkers. Those skills are verbal reasoning, argument analysis, thinking to test hypotheses, understanding and applying likelihood, uncertainty and probability, decision making and problem solving (Halpern, 1998). The dispositions Halpern described are a willingness to engage and persist with complex tasks, habitually planning and resisting impulsive actions, flexibility or open-mindedness, a willingness to self-correct and abandon non-productive strategies and an awareness of the social context for thoughts to become actions.

In education research there is often an emphasis on critical thinking as a skill set (Bailin, 2002) or putting critical thought into tangible action (Barnett, 1997). Dressel and Mayhew (1954) suggested it is educationally useful to define critical thinking as the sum of specific behaviours which could be observed from student acts. They describe these critical thinking abilities as identifying central issues, recognising underlying assumptions, evaluating evidence or authority, and drawing warranted conclusions. Bailin (2002) raises the point that from a pedagogical perspective many of the skills or dispositions commonly used to define critical thinking are difficult to observe and, therefore, difficult to assess. Consequently, Bailin suggests that the concept of critical thinking should explicitly focus on adherence to criteria and standards to reflect 'good' critical thinking (Bailin, 2002, p. 368).

It appears that there are several definitions of critical thinking of equally valuable meaning (Moore, 2013). There is agreement across much of the field that meta-cognitive skills, such as self-evaluation, are essential to a well-rounded process of critical thinking (Glaser, 1984; Kuhn, 1999; Pithers and Soden, 2000). There are key themes such as ‘critical thinking: as judgement, as scepticism, as originality, as sensitive reading; or as rationality’ which can be identified across the literature. In the context of developing an individuals’ critical thinking it is important that these themes take the form of observable behaviours.

To summarise, Table 1. 1 illustrates the commonalities and divergences of how the disciplines of philosophy, cognitive psychology and education research define critical thinking with respect to the skills they use to describe it. Broadly speaking there is a consensus amongst these groups with the exception of cognitive psychologists’ application of evaluation. The cognitive psychologists add deductive logic, decision making and problem solving to their definition. Self-regulation or meta-cognitive skills are shown at the bottom of the table and underpin the definitions drawn from all three disciplines.

Table 1. 1 Categorization of critical thinking skills as defined by the philosophy, cognitive psychology and education research disciplines

Philosophers^a	Cognitive Psychologists	Education Researchers^d
Interpretation	Apply likelihoods ^c	Recognising underlying assumptions
Analysis	Argument analysis ^c	Identifying central issues
Evaluation		Evaluating evidence or authority
Inference	Statistical reasoning ^b	
Explanation	Verbal reasoning ^b	Drawing warranted conclusions
	Methodological reasoning ^b	
	Conditional reasoning ^b	
	Decision making ^c	
	Problem solving ^c	
	Thinking to test hypotheses ^c	
Self-regulation ^a /meta-cognitive skills ^e		

^a(Facione, 1990) , ^b(Lehman and Nisbett, 1990; Nisbett *et al.*, 1987), ^c(Halpern, 1993), ^d(Dressel and Mayhew, 1954), ^e(Glaser, 1984; Kuhn, 1999)

1.2 How is Critical Thinking Taught?

There are two opposing views regarding the teaching of critical thinking and the role discipline specific knowledge plays in its development. To understand these opposing viewpoints it is useful to briefly explore the history of teaching critical thinking and how these

perspectives have evolved. Methods of teaching critical thinking began in ancient Greece with Socrates and Plato. These scholars referred to critical thinking as 'formal discipline' (Mann, 1979). Formal discipline was the concept that the laws of logic and reason were best developed through the study of mathematics and geometry. Formal discipline focused on the formative education of the mind rather than instruction for the purpose of imparting knowledge or information. Plato believed that logic and reasoning were transferable skills and that a curriculum focused on topics which had no utilitarian use, such as pure mathematics or geometry, were most effective in preparing the mind for operations in other unrelated areas (Mann, 1979).

The Romans valued many of the pedagogical ideas developed by the Greeks but focused on the practical applications formal discipline could provide. The abstraction of reasoning and logic were most valued by upper class citizens who could leverage these skills in the political arena. In the lower levels of Roman society, cognitive training was focused on preparing students for success in their careers. Quintilian, one of the most notable Roman educators, considered the study of language as an important tool for developing his pupils. He valued the contribution that the study of mathematics made to his students deductive reasoning skills, but felt the memorisation of poetry and the study of language and grammar were crucial in sharpening wits and modulating thought (Mann, 1979).

In medieval times there was a strong emphasis on logic and syllogisms. Around the 15th century, the study of Latin and Greek was added to the concept of formal discipline. The notion of formal discipline as a means to teach reasoning and logic were strongly advocated well into the 19th century. Disregarding Plato, John Locke is considered the contemporary father of formal discipline. Locke believed strongly in the capacity of formal discipline, in subjects such as mathematics, to be able to develop cognitive training and to be able to transfer that training to other circumstances. He is quoted as saying 'Nothing does this (cognitive training) better than mathematics ... not so much to make them mathematicians, as to make them reasonable creatures' (Mann, 1979, p. 138). The sentiment of teaching a field strictly for developing cognitive discipline is also echoed by later scholars with respect to

language; 'The acquisition of language (Latin in this case) is educationally of no importance; what is important is the process of acquiring it' (Mann, 1979, p. 149). These viewpoints went largely unchallenged until the 20th century.

At the start of the 20th century a prominent series of research papers were published by Thorndike and co-workers, which critiqued formal discipline (Thorndike and Woodworth, 1901a, 1901b, 1901c). Their participants were exposed to a series of exercises focused on estimating the area of various shapes and/or recognising spelling errors or strange fonts in a body of text. The researchers asked the participants to estimate the geometries of different shapes and sizes. The participants were then taught skills in estimating the dimensions of rectangles of specific sizes, but when asked to repeat the initial task they still exhibited the same degree of difficulty with the task as they did prior to the intervention. When the participants practiced marking words with the letters 'e' and 's' until a considerable improvement was made in speed and accuracy, there were insignificant improvements in the participants' ability to mark misspelt words or other combinations of letters before and after the training. Thorndike and his colleagues thus developed the alternate view of discipline specificity and questioned the capacity for subjects such as Latin or pure mathematics to teach reasoning and logic.

Inhelder and Piaget (1958) further disputed the value of formal discipline to develop abstract logic and reasoning. They believed that the ability of an individual to transfer logic and reasoning skills into other disciplines was limited. They believed that without the concrete knowledge of a specific discipline, one could not adequately apply reason and logic in formal operations.

The limitations of formal discipline were illustrated by Wason's selection test (Wason, 1966). The logic used in his test followed the typical "if p , then q " format requiring the participant to prove that p is in fact q and also that not q is not p . He found that in various populations participants struggled to answer problems in his abstract test, thus supporting the importance of discipline specificity as postulated by Thorndike.

Some resistance to the ideas of the early 20th century emerged in the 1950s in what came to be termed informal logic (Johnson, Blair and Hoaglund, 1996). Informal logic gained academic credence in the 1970s as it challenged the previous ideas of logic being related purely to deduction or inference, and that there were in fact theories of argumentation and logical fallacies. These theories began to be taught at universities as standalone courses free from any context in efforts to teach the structure of arguments and recognition of fallacies using abstract theories and symbolism.

McPeak challenged the growing popularity of informal logic in his book 'Critical Thinking and Education' (McPeak, 1981). He stated that thinking is never without context and thus courses designed to teach informal logic in an abstract environment provide no educational benefit to the student's capacity to think critically (McPeak, 1990).

In later years cognitive psychology lent evidence to the argument that critical thinking could be developed within a specific discipline and those reasoning skills were, at least to some degree, transferable to situations encountered in daily life. Lehman, Lempert and Nisbett (1988) conducted cross sectional and longitudinal studies with graduate students from the schools of medicine, psychology, law, social science and chemistry based on their performance in two reasoning tests. These tests were administered at the beginning and conclusion of postgraduate studies in the form of a pre- and a post-test. Items contained on the test were aimed at measuring conditional, statistical, methodological and verbal reasoning. *Conditional reasoning* was defined as schemas where permission or an obligation for an event to occur is either satisfied or violated in the general form of premises such as 'if p , then q '. *Statistical reasoning* was the term used to describe the inferential reasoning applied when assessing the value of a sample relative to the population. *Methodological reasoning* was that which is employed when making judgements regarding sample and personal bias, errors in self-selected participants and the need for control groups. Finally *verbal reasoning*, was believed to improve with maturity and was used as a control. Lehman *et al.* (1988) concluded that the probabilistic sciences of medicine and social psychology exposed students to statistical and methodological reasoning both in scientific and daily life problems, whilst law

and the non-probabilistic sciences of chemistry and natural psychology did not. Similarly, they concluded that social psychology, medicine and law students better developed their conditional reasoning as a result of being taught pragmatic reasoning schemas akin to the permission and obligation schemas of conditional reasoning.

Lehman and Nisbett (1990) established a similar link between reasoning and training with undergraduate students. A longitudinal study of undergraduate students from a range of disciplines such as natural sciences, humanities, social sciences and psychology was conducted using a pre-test which was administered in the first year of the participants' studies and a post-test during the fourth year of their studies. Much like the researchers' studies with postgraduate students, tests were comprised of items which tested statistical, methodological, conditional and verbal reasoning. The undergraduate students also developed specific forms of reasoning dependent on the discipline studied. Psychology and social science students made significantly large gains in statistical and methodological reasoning, while natural science and humanities students made small but significant gains. Natural science and humanities students showed significant improvement in conditional reasoning while psychology and social science demonstrated no gains. The findings obtained from the studies with the undergraduates and postgraduates led to the conclusion that the development of specific reasoning skills was related to the discipline undertaken by the students.

McMillan (1987) conducted a review of 27 empirical studies performed at higher education institutions where critical thinking was taught, either in standalone courses or integrated as part of discipline-specific courses, such as science. The review found that standalone and integrated courses were equally successful in developing critical thinking, provided critical thinking developmental goals were made explicit to the students. The review also suggested that the development of critical thinking was most effective when its principles were taught across a variety of discipline areas so as to make knowledge retrieval easier.

Ennis (1989) suggested that there are a range of approaches through which critical thinking can be taught: *General*, where critical thinking is taught separate from content or 'discipline'; *Infusion*, where the subject matter is covered in great depth and teaching of critical

thinking is explicit; *Immersion*, where the subject matter is covered in great depth but critical thinking goals are implicit; and *Mixed*, a combination of the general approach with either the infusion or immersion approach. This model allowed Ennis to suggest three types of discipline specificity with respect to developing critical thinking: Domain Specificity, Epistemological Subject Specificity and Conceptual Subject Specificity (Ennis, 1990). He provided the principles for each type of specificity derived from evidence from cognitive psychology. *Domain specificity* is the view that critical thinking comprises of three principles: background knowledge, that critical thinking skills are unlikely to be transferred from one domain to another without explicit instruction, and that critical thinking is unlikely to be learnt in a general critical thinking course. *Epistemological Subject Specificity* is formed around three principles too: background knowledge is required to make justified critical thinking judgements, critical thinking varies from field to field, and a full understanding of a field requires the ability to think critically within that field. *Conceptual Subject Specificity* focuses on the principle that general critical thinking skills do not exist and teaching critical thinking outside of a subject matter area is redundant.

Ennis (1990) arrived at a pragmatic view to concede that the best critical thinking occurs within one's area of expertise, or domain specificity, but that critical thinking can still be effectively developed with or without discipline specific knowledge (Ennis, 1990; McMillan, 1987). Despite acknowledging Ennis's viewpoint, many scholars still remain entrenched in the debate regarding the role discipline specific knowledge has in the development of critical thinking. For example, Moore (2011) rejected the use of critical thinking as a catch-all term to describe a range of cognitive skills, believing that to teach critical thinking as a set of generalisable skills is insufficient in providing students with an adequate foundation for the breadth of problems they will encounter throughout their studies. To emphasise his point, (Moore, 2013) conducted interviews with colleagues in history, literature and philosophy, discussing the nature of critical thinking in the academics' respective disciplines. From these interviews he suggested that all critical thinking requires the 'rendering of judgement'. However, the nature of judgement, be it constructing, evaluating or understanding arguments,

varied between disciplines and was dependent on the 'object of inquiry'. To this end, Moore is opposed to the generalist approach of standalone critical thinking courses, instead advocating for students to develop the flexibility of thought to engage in many critical modes. Moore terms this flexibility of thought as 'metacritique' (Moore, 2011, p. 262), and prefers the development of critical thinking to occur as part of discipline-specific studies, analogous to Ennis's *infusion* approach. Davies (2013) accepted that critical thinking skills share fundamentals at the basis of all disciplines, and that there can be a need to accommodate the discipline-specific needs 'higher up' in tertiary education via the *infusion* approach. However, Davies considers the specifist approach to developing critical thinking 'dangerous and wrong-headed' (Davies, 2013, p. 543), citing government reports and primary literature which demonstrates tertiary students' inability to identify elements of arguments, and championing the need for standalone critical thinking courses.

Figure 1. 1 summarises the evolution of how critical thinking is taught and the role of discipline specific knowledge. In summary, how critical thinking is taught dates back to the times of Socrates, Plato and Quintilian. These scholars taught critical thinking as formal discipline through the study of mathematics, geometry and language. They believed the mental acumen developed from these studies allowed their students to be better thinkers in aspects of life unrelated to mathematics and language such as politics or military leadership. The foundations set by these great scholars suggested that good critical thinking could be taught independent of context, which would later become known as the generalist perspective on the role of discipline knowledge in developing critical thinking. The generalist perspective went unchallenged until the first half of the 20th century through the work of psychologists Thorndike, Piaget and Wason. These psychologist developed empirical evidence and theories to suggest that critical thinking was entirely dependent on discipline knowledge in what became known as the specifists perspective. Over time the pedagogical approach of teaching critical thinking in the abstract form of informal logic emerged, and academics such as McPeak and Ennis argued over whether critical thinking should be taught dependently or independently of context. At the turn of the 21st century scholars had become less extreme in their views of the role of discipline

specific knowledge when teaching critical thinking. While academics still favour either a specifists or generalist view, both sides of the argument acknowledge that the best critical thinking is achieved with the foundation of discipline specific knowledge, and the critical thinking skills developed in one context have some degree of utility in other contexts (Davies, 2013; Moore, 2013).

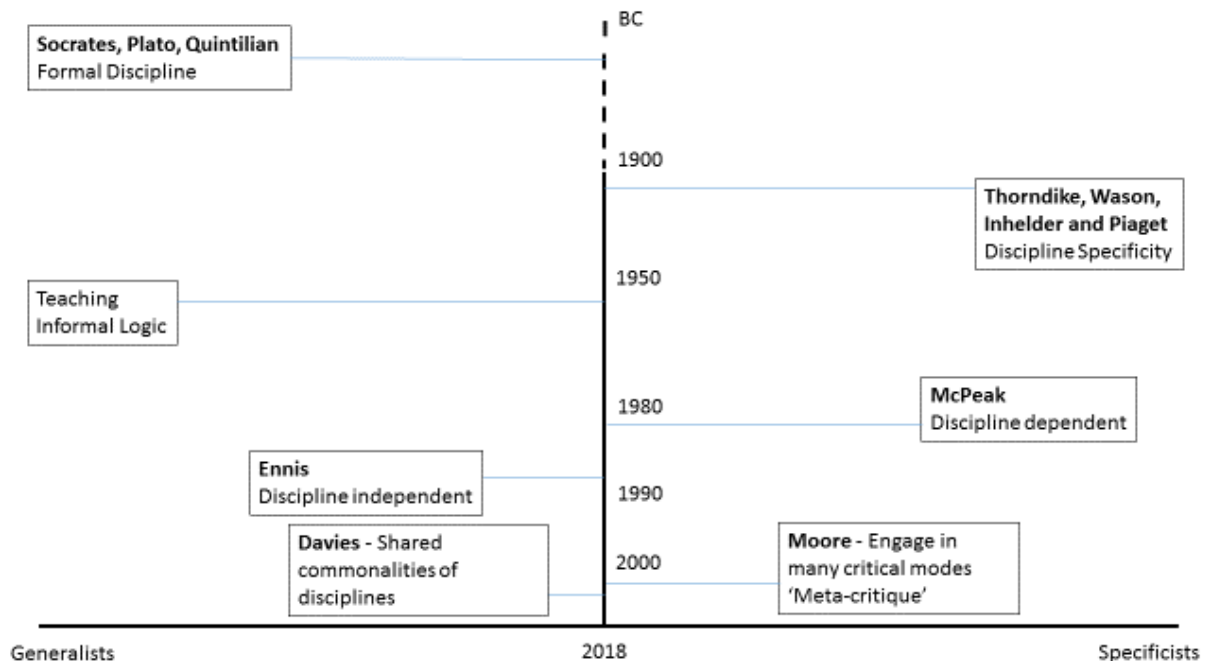


Figure 1. 1 Timeline highlighting the evolution of the role of discipline specific knowledge in critical thinking. Distances on the horizontal axis indicates proximity from generalist or specifist views

1.2.1 Recent Approaches Used to Develop Critical Thinking

Using Ennis's (1989) terminology, recent teaching pedagogies range from the *general* approach of teaching critical thinking through explicitly designed courses (Jacob, 2004), to the *immersion* approach of embedding critical thinking material implicitly within a discipline course (Davies, 2006). Many reviews suggest that the best teaching outcomes occur when the development of critical thinking is made explicit to the student and that this development occurs over the entire duration of their studies rather than through a single course or semester (Abrami *et al.*, 2008; Tiruneh *et al.*, 2014; Tsui, 2002).

Pedagogical approaches to developing critical thinking at higher education institutions range from writing exercises (Martineau and Boisvert, 2011; Oliver-Hoyo, 2003; Stephenson

and Sadler-Mcknight, 2016), inquiry-based projects (Gupta, Burke, Mehta and Greenbowe, 2015), flipped lectures (Flynn, 2011) and open-ended practicals (Klein and Carney, 2014) to gamification (Henderson, 2010), and work integrated learning (WIL) (Edwards, Perkins, Pearce and Hong, 2015). Researchers have demonstrated the benefits of developing critical thinking skills across all first, second and third year programs of an undergraduate degree (Iwaoka, Li and Rhee, 2010; Phillips and Bond, 2004). The study by Phillips and Bond (2004) highlighted an opportunity to embed critical thinking skills within the entire undergraduate chemistry syllabus, and design interventions whereby students were exposed to a high number of ill-defined problems throughout their studies. Phillips and Bond (2004) indicated that such interventions help develop a culture of inquiry, and better prepare students for employment.

Studies have demonstrate the outcomes of their teaching interventions via validated commercially available critical thinking tests (Abrami *et al.*, 2015; Abrami *et al.*, 2008; Carter, Creedy and Sidebotham, 2015; Tiruneh *et al.*, 2014). There are arguments against the generalisability of these commercially available tests, partly because many academics believe assessments need to closely align with the intervention(s) (Ennis, 1993). They believe a more accurate representation of student ability is obtained when a critical thinking assessment is related to a students' discipline, as they attach greater significance to the assessment (Halpern, 1998).

1.3 The Assessment of Critical Thinking

As constructive alignment suggests (Biggs, 2012), good teaching requires assessment which effectively assesses the outcomes an educator wants students to be able to demonstrate. Traditional methods of assessment, such as summative examinations and the outputs from expository laboratories, such as reports, are limited in their ability to evaluate students' critical thinking skills. A study of over 80 first and third year chemistry examinations papers from the UK and Australia demonstrated that 90% of problems within the examinations were algorithmic in nature. That is, the input data is given, the methods are familiar and the outcomes are closed responses (Bennett, 2008). Domin (1999) performed a content analysis

of ten general chemistry laboratory manuals and found that the laboratory activities required students to use only the lower levels of Bloom's taxonomy (Ramirez, 2017) rather than higher order cognitive skills such as analysis, synthesis and evaluation. The evaluations conducted by Bennett (2008) and Domin (1999) of traditional methods of assessing chemistry via examinations and student reports from laboratories suggested that students became proficient at these closed problem solving skills. While at the same time the evaluation of students' ability to solve problems with incomplete data, using unfamiliar methods and open-ended outcomes remained under-assessed (Ramsden, 1992). Examinations and laboratory reports provide students with opportunities to develop lower levels of the Bloom's taxonomy (Ramirez, 2017), however these forms of assessment do not provide students with the opportunity to develop higher level skills, such as evaluation and judgement which are considered to be signs of good critical thinking (Facione, 1990).

1.3.1 Approaches to Assessing Critical Thinking in Higher Education Chemistry

In recent years educators have attempted to demonstrate gains in the development of students' critical thinking through a variety of teaching interventions and assessment tools. Flynn (2011) used clickers in large flipped classrooms of 200 to 500 chemistry students to promote problem solving and critical thinking skills by providing students with real examples of complex organic molecules which contained several possible retrosynthetic approaches. Students were required to suggest their preferred retrosynthetic pathway and the reagents that would achieve their approach. While the kind of thinking was somewhat declarative in nature and there was no empirical data to suggest the development of critical thinking, the authors considered the complexity of the problems and variety of approaches students could take was reflective of critical thinking.

Henderson (2010) implemented a chemical instrumentation game designed to take place in a tutorial setting. Students took on the role of representatives from an instrumentation company and were required to make a proposal to their peers (playing the role of vice-presidents) and the tutor (the CEO) regarding the usefulness of an item of analytical equipment, such as a mass spectrometer. The proposals took the form of a ten-minute

presentation for which students were required to have consulted the academic literature and present responses to prompts such as 'how is the technology novel?' and discuss 'the significance of selected journal articles related to the proposal'. In addition, the students were required to write an annotated bibliography of at least five articles directly related to their proposal. A novel form of peer evaluation was used whereby the students decided how much they wanted to invest in the technique based on the quality of the presentations. The authors noted that this created a very competitive environment and led to students being very critical of their peers using thorough questioning. The author observed that students who typically excelled at examinations found the task challenging and attributed this to the exercise requiring skills such as critical thinking.

Klein and Carney (2014) designed a capstone project running over two units which they believed developed students' critical thinking. The first unit was run in the third year of a chemistry course with approximately 16 students. These students were required to write a literature review of 20-25 pages in length and deliver an 18-minute presentation on a research topic of their choice under the guidance of a faculty member. This work served as the introduction and background to the research project that the students subsequently undertook in their fourth year of chemistry study. The authors collected qualitative evidence from a graduate exit survey which consisted of Likert and open-ended responses. The students expressed that they appreciated that the capstone unit provided them with the opportunity to focus on reading, analysing and summarising primary literature, and the written and oral communication assignments, which the authors considered elements of critical thinking.

Gupta *et al.* (2015) were interested in the use the scientific writing heuristic (SWH) by (Keys, Hand, Prain and Collins, 1999) to assist with students' scientific report writing in order to develop their critical thinking skills. The researchers compared the scientific reports of first year chemistry students who underwent the SWH intervention with the reports of first year and fourth year chemistry students who did not undergo the intervention using a marking rubric which accompanies the SWH. The rubric contains scales to measure cognitive traits related to critical thinking such as organisation, comprehension and evaluation. The analysis of the

student reports provided statistically significant data indicating students who underwent the SWH intervention performed better in the areas of evaluation, analysis and comprehension compared to students who did not undergo the SWH intervention. From the findings the authors concluded that the SWH was suitable for the development and assessment of student critical thinking.

Weaver, Samoshin, Lewis and Gainer (2016) constructed inquiry-based laboratory exercises during which 90 organic chemistry students were required to synthesise molecules of known pharmaceutical interest. The students were required to search the literature and propose their synthetic route, generate and analyse these molecules in laboratories and report their findings in the form of a manuscript to the academic journal *Organic Letters*. The authors collected student perspectives, which indicated students believed they developed their problem solving and critical thinking skills through the exercise.

1.4 Commonly Used Critical Thinking Tests and Teaching Tools

Many researchers use validated commercially available tests such as the Californian Critical Thinking Skills Test (CCTST), the Cornell Critical Thinking Test (CCTT), the Critical Thinking Assessment Task (CAT) or the Watson-Glaser Critical Thinking Appraisal (WGCTA) to evaluate student development of critical thinking and/or teaching interventions designed to develop critical thinking. Several recent reviews have shown a variety of applications of these commercially available tests, ranging from testing pre- and post-teaching interventions, to cross-sectional or longitudinal studies (Behar-Horenstein and Niu, 2011; Carter *et al.*, 2015; Huber and Kuncel, 2016; O'Hare and McGuinness, 2015).

Examples of these tests being utilised in chemistry education research include authors such as Chase *et al.* (2017), who used the CAT in a pre- and post-study to demonstrate gains in critical thinking upon delivery of an authentic research laboratory experience to 86 undergraduates in an advanced general chemistry course. Quitadamo *et al.* (2011) compared the critical thinking skills of first year biology and chemistry students upon entering university using the CCTST. The authors showed that chemistry students typically performed better in

areas of inference and evaluation relative to biology students. They also identified that completion of high school physics was a significant variable in determining performance on the CCTST, and that a larger proportion of chemistry students had undertaken high school physics relative to biology students. The authors suggested that the mathematical nature of physics develops logical thinking, or what Lehman and Nisbett (1990) would consider conditional reasoning.

Since the early 1900s critical thinking tests have been developed for a variety of reasons. They are typically used to evaluate potential employees or entrants into specific university courses (Hambur, Rowe and Luc, 2002; Insight Assessment, 2013; Pearson, 2015), or used by education researchers for a host of pedagogical reasons such as the longitudinal develop of students or evaluation of teaching interventions (Gadzella, Hogan, Masten and Stacks, 2006; Hassan and Madhum, 2007). Tests may be administered to participants from pre-school, secondary school to higher education or at the professional level (Carter *et al.*, 2015; Drennan, 2010; Pithers and Soden, 1999). For the purposes of this research the discussion will focus primarily on the tests and teaching tools used within the higher education setting. Ennis and Chatten (2015) categorises these tests into three groups; commercially available tests which have undergone large bodies of reliability and validity testing, tests developed by governmental organisations, and tests developed by researchers specifically to answer their particular research question or for the development of critical thinking.

The following tests were originally developed by notable academics in the fields of education and/or psychology. They have typically undergone reliability and validity studies with thousands of participants and over the years have been revised and updated. These tests were eventually commercialised to target customers such as employers or universities. Whilst this list may not be exhaustive (see Table 1. 2), it highlights the tests most commonly reported in the literature.

Table 1. 2 Summary of commonly used commercially available critical thinking tests

Test	Question structure	Critical thinking skills assessed
CCTST (Insight Assessment, 2013)	40 item multiple choice questions	analysis, evaluation, inference, deduction and induction
WCGTA (TalentLens, 2017)	85 item multiple choice questions	inference, deduction, drawing conclusions, making assumptions and assessing arguments
WGCTA-S (Pearson, 2015)	40 item multiple choice questions	inference, deduction, drawing conclusions, making assumptions and assessing arguments
CCTT-Z (The Critical Thinking Co., 2017)	52 item multiple choice questions	induction, deduction, credibility, identification of assumption, semantics, definition and prediction in planning experiments
EWCTET (Ennis and Weir, 1985a)	Eight paragraphs which are letters containing errors in critical thinking and essay in response to these paragraphs.	understanding the point, seeing reasons and assumptions, stating one's point, offering good reasons, seeing other possibilities and responding appropriately and/or avoiding poor argument structure
HCTA (Halpern, 2016)	20 scenarios or passages followed by a combination of 25 multiple choice, ranking or rating alternatives and 25 short answer responses	reasoning, argument analysis, hypothesis testing, likelihood and uncertainty analysis, decision making and problem solving

The California Critical Thinking Skills Test (CCTST) was first developed by Facione. It is a 40 item multiple choice test designed to assess analysis, evaluation, inference, deduction and induction skills (Insight Assessment, 2013). Respondents are presented with several diagrams or short passages and select the most appropriate response from a selection of multiple choice answers for each question. Participants obtain a score out of 40, which is compared to the norms established in test manual. The test is typically not timed and exists in two forms to allow pre- and post-testing. It is available in paper or digital formats. When searching the recent literature for assessment of critical thinking this test is the most prevalent form of evaluation.

Accompanying the CCTST is the California Critical Thinking Disposition Inventory (CCTDI) (Insight Assessment, 2017). While the CCTST aims to evaluate a participant's critical thinking skill the CCTDI aims to evaluate the likelihood with which an individual will engage in behaviour which puts those critical thinking skills into tangible action. This test is unique in that it is explicitly an assessment of critical thinking behaviour rather than critical thinking skill. Much like the CCTST the respondents are presented with 40 statements followed by a series of multiple choice responses to the scenarios. These are designed to highlight an individual's propensity to exhibit behaviours such as truth seeking, open-mindedness, analyticity, systematicity, confidence, inquisitiveness and maturity of judgement (Facione *et al.*, 1995). As with the CCTST it is generally untimed, available in two formats, and paper or digital versions. Whilst not essential to the measurement of critical thinking, it has been highlighted previously that good critical thinking requires not only the skills but also the motivation to use those skills (Bailin, 2002). Some researchers use the CCTST and CCTDI in tandem to obtain a rounded view of critical thinking whilst others use the CCTDI in conjunction with other behavioural measurement methods.

The Watson-Glaser Critical Thinking Appraisal (WCGTA) (AssessmentDay Ltd, 2015) is the second most commonly used test to evaluate critical thinking. Watson and his then PhD student, Glaser, wrote the first version in 1920. The test has the longest history of reliability and validity, and undergone numerous revisions since its inception. The test is designed to evaluate the critical thinking skills of inference, deduction, drawing conclusions, making assumptions and assessing arguments (TalentLens, 2011). This test comprises of 85 multiple choice questions and, with the exception of the inference section, there are several short statements followed by prompts containing dual choice responses of A or B. In the case of the inference section the participants are provided with a paragraph of text followed by an inference and five multiple choice options from which to decide the quality of that assumption with respect to the paragraph of text. Participants obtain a score out of 80 and a score for each section. However, the authors of the test discourage the use of the section scores, as the authors concede there are insufficient items in a given section to obtain any meaningful insights

(TalentLens, 2011). The test provides thorough instructions and examples for each section, instructions to the facilitator and typically is not timed. This test is available in an A or B version so as to facilitate pre- and post-measurements. The authors of the WCGTA have developed the WCGTA short form (WGCTA-S) (Pearson, 2015). This is a condensed version of the 85 item test containing all the same sections, instructions and formatting. It uses just 40 questions but has also undergone rigorous validity and reliability testing (Watson and Glaser, 2006). The WGCTA-S is available in two forms (A and B). All versions of these tests are available in paper format or for online delivery.

The Cornell Critical Thinking Test Level Z (CCTT-Z) is another multiple choice test. Co-authored by Robert Ennis and Jason Millman (The Critical Thinking Co., 2017), it is probably the third most common test used by researchers when evaluating their students' critical thinking. It is a 52 item test measuring areas of induction, deduction, credibility, identification of assumption, semantics, definition and prediction in planning experiments. It can be administered in a 50-minute time slot or untimed. To aid in its administration a manual is available with norms, validity, reliability and item analysis. There are many iterations of the CCTT, however the CCTT-Z specifically targets university students.

The Ennis-Weir Critical Thinking Essay Test (EWCTET) was written by Robert Ennis and Eric Weir (Ennis and Weir, 1985b). It covers critical thinking competencies such as 'understanding the point', 'seeing reasons and assumptions', 'stating one's point', 'offering good reasons', 'seeing other possibilities' and 'responding appropriately and/or avoiding poor argument structure', for example credibility. This test is unique in that, rather than presenting a series of multiple choice questions, participants are provided with eight paragraphs which are letters to the editor of a fictional newspaper. These letters exhibit errors in critical thinking and the participants are asked to write an essay in response to these paragraphs. The test moves away from the deductive or formal logic used in other tests and attempts to frame critical thinking in a real world context and, to some degree, allows the participant to exhibit critical thinking dispositions. Another distinction of this test is that graders are provided with marking criteria and scoring instructions with which graders are encouraged to use their own judgement

when applying the marking scheme. Typically this test is administered in the space of 40 minutes, with an additional 10 minutes of reading time.

Developed by Diane Halpern, the Halpern Critical Thinking Assessment (HCTA) (Halpern, 2016) claims to be the only general domain critical thinking test to use a format combining 'forced choice' (which include multiple choice, ranking or rating alternatives) and 'constructed response' (more akin to short answer responses). The test uses a combination of 25 'forced responses' and 25 'constructed responses'. The HCTA focuses on reasoning, argument analysis, hypothesis testing, likelihood and uncertainty analysis, decision making and problem solving. There are 20 scenarios or passages in the areas of everyday education, health, politics and social issues, which are then followed by a series of questions. The test is available as either a paper or digital based format and administrators have access to a breadth of norms to compare the data. Halpern claims that written responses and recall responses elicit different types of thinking. Therefore, while essay style responses may underestimate the critical thinking skills of poor writers, essay style responses may somewhat capture the dispositions of critical thinking.

Public bodies often develop valid and reliable assessments to serve a variety of purposes. For example, a test may be used for entry into specific university degrees or used to measure the attributes of exiting graduates (ACT, 2017; Hambur *et al.*, 2002). These assessments consider a breadth of skills and generally critical thinking forms one of these sections. Researchers use the critical thinking component of one of these tests as a form of convergent validity with either one of the commercially available tests described above or with a test of their own design (Blattner and Frazier, 2002; Stein, Haynes, Redding, Ennis and Cecil, 2007; Tiruneh, De Cock, Weldeslassie, Elen and Janssen, 2016). This approach is somewhat infrequent. However, use of the tests described below appear in the literature with enough frequency to warrant a brief discussion.

The Collegiate Assessment of Academic Proficiency (CAAP) is a tool developed by the American College Testing (ACT) program. The critical thinking module contains 32 multiple choice questions which assess analysis, evaluation and the extension of arguments (ACT,

2017). The test consists of four passages comprised of a variety of case studies, experimental data and debates on which the responses to the multiple choice questions are based (Terenzini, Springer, Pascarella and Nora, 1995).

The graduate skills attributes (GSA) test is a competency test designed by the Australian Council for Education Research (ACER) to evaluate graduates upon exiting university. The critical thinking component consists of 30 multiple choice questions which follow a series of text passages. The GSA assesses comprehension, analysis, inference, synthesis and evaluation. However, ACER concede that the 30 questions do not cover the breadth of skills used in critical thinking (Hambur *et al.*, 2002). All items use what the authors of the GSA refer to as 'everyday contexts' as they acknowledge the ability to think critically is dependent on familiarity with the context. The GSA, however, is no longer available from ACER (Badcock, Pattison and Harris, 2010).

There are also a number of critical thinking tests developed by researchers for their own specific needs. These are often discipline specific and are not validated to the same extent as the commercially available tests. This lack of validation may be due to the fact that the tests were only ever intended for internal use. Whilst there are critical thinking tests for almost all disciplines taught at higher education institutions, from nursing (Carter *et al.*, 2015; Daly, 2001) to economics (Ennis and Chatten, 2015), the following section highlights a key selection of the tests or teaching tools specifically developed by science and chemistry educators and education researchers.

The Lawson Classroom Test of Scientific Reasoning (LCTSR) (Lawson, 2000) is based around assessing what Piaget referred to as the participants' developmental level relative to scientific concepts, such as conservation of weight and volume, proportional reasoning, probabilistic reasoning, control of variables and hypothetico-deductive reasoning (Carmel and Yeziarski, 2013). The test is divided into 12 items, each one containing two multiple choice questions. The first parts of each question asks the participant to make a selection influenced heavily by the deductions of observations provided in the question. The second part of each

question then requires the participant to select the reasoning for their choice in the first part of the question.

The chemistry concept reasoning test was designed to evaluate the conceptual understanding and resulting critical thinking of first year chemistry (general chemistry in the US) students (Cloonan and Hutchinson, 2011). There are two versions of the test to allow for pre- and post-testing. The test was designed to assess the conceptual understanding of atomic theory, atomic structure, chemical bonding, chemical equilibrium, chemical kinetics, chemical reactions, kinetic molecular theory, phase equilibrium, thermochemistry and thermodynamics. It is a multiple choice test that provides participants with a series of statements and the participants are required to identify which of these statements are conceptually correct. Then the participant must select a response which describes how the various statements relate to each other to construct a logical argument. This test is dependent on declarative knowledge and determining logical reasoning which is deductive in nature.

The Critical Thinking Assessment Task (CAT) is a critical thinking test oriented towards STEM disciplines developed by the Tennessee Technological University (Stein and Haynes, 2011). It uses 15 short answer responses to assess inference, assessing arguments, correlation versus causation, analysing, interpreting and evaluating information, communicating an argument and algorithmic problem solving. The questions are designed to reflect real world problems with a focus on engaging with new information. The authors acknowledge that scoring reliability of essay type question can be problematic with respect to consistency between graders, however, the researchers believe short answer questions provide better insight into the thought processes of the participant compared to multiple choice questions.

1.4.1 Chemistry Specific Tests and Teaching Tools

Jacob (2004) developed a short questionnaire to evaluate chemistry students' critical thinking ability. Students were presented with six questions each consisting of a statement requiring an understanding of declarative chemical knowledge. The participants were asked to highlight the chemicals in the question, the experimental observations and the conclusions

drawn in each question. From the words highlighted, the students then needed to select whether the conclusion was valid, possible or invalid and provide a short statement to explain their reasoning. This tool was developed as the researcher was interested in the effect of explicitly teaching logic rules to chemistry students alongside their discipline-specific content, which Ennis (1989) would refer to as an immersion approach.

Kogut (1993) developed a series of chemistry critical thinking exercises to assist in developing the critical thinking of introductory chemistry students. These exercises were designed to generate a dialogue between the students and academic within a lecture setting. The exercises were intended to prompt students to note observations and underlying assumptions of chemical phenomena then develop hypotheses, and experimental designs with which to test these hypotheses. However, understanding the observations and underlying assumptions was dependent on declarative chemical knowledge such as trends in the periodic table or the ideal gas law. The author observed that the teaching intervention increased students' evaluative and sceptical approach to science, and students engaged in greater independent learning. Although the author acknowledges critical thinking growth is difficult to measure, their experience with students' responses to questions, recognition of assumptions and problem solving in relation to these exercises provided qualitative indicators of improvements in critical thinking.

Garratt, Overton and Threlfall (1999) developed an entire book dedicated to developing chemistry critical thinking, titled 'A Question of Chemistry'. In writing this book the authors took the view that thinking critically in chemistry draws on the generic skills of critical thinking and what they call 'an ethos of a particular scientific method' (Garratt, Overton, Tomlinson and Clow, 2000, p. 153). To this end their work specifically challenges students in the areas of analysis and evaluation of arguments, making judgements, retrieving information and experimentation. The approach to delivering these questions ranged from basic multiple choice questions, to rearranging statements to generate a cohesive argument, or open-ended responses. The common approach amongst these styles of questioning was that students were organised into groups of approximately three to five and asked to discuss what they

believed to be their 'preferred response'. The statements preceding the questions are very discipline specific and the authors acknowledge they are inaccessible to a lay person. Overall the chemistry context is used because 'it adds to the students' motivation if they can see the exercises are firmly rooted in, and therefore relevant to, their chosen discipline' (Garraatt *et al.*, 2000, p. 166).

1.5 Review of Critical Thinking Assessment Tools

Several reviews and meta-analyses have been conducted on published studies which utilise previously mentioned critical thinking tests (Behar-Horenstein and Niu, 2011; Carter *et al.*, 2015; Huber and Kuncel, 2016; O'Hare and McGuinness, 2015). These articles investigated the various tests used, their research methodologies, applications, reliability and validity, and the outcomes of studies that used these various tests.

Behar-Horenstein and Niu (2011) conducted a review of the literature between 1994 and 2009 to find 42 empirical studies using critical thinking assessment tools. They found the most commonly used commercial tests were the CCTST, the CCTT-Z, and the WGCTA. They found 45% used the CCTST, 45% used the WGCTA, and the remaining 4% used the CCTT-Z. Carter *et al.* (2015) conducted a thorough review of the literature between 2001 and 2014 to find that 12 studies utilised the CCTDI, seven used the CCTST, and three used the WGCTA. In their meta-analysis of 71 reports of pre/post studies from 1966 to 2012, Huber and Kuncel (2016) found that 55 studies used the WGCTA, 25 studies used the CCTST, 11 used the CCTDI and 12 used the CCTT-Z. At face value, Huber and Kuncel's (2016) study suggests the WGCTA is used more frequently. However, the CCTST was developed much later than the WGCTA and the recent increase in the popularity of the CCTST is reflected in the data of Behar-Horenstein and Niu (2011) and Carter *et al.* (2015).

Typically, in the empirical research, the tests were administered to address questions regarding the development critical thinking over time or the effect of a teaching intervention. To obtain these results, researchers administered pre- and post-tests or conducted cross-sectional studies. Huber and Kuncel's (2016) analysis found 16,185 longitudinal studies and 9,392 cross sectional studies. Behar-Horenstein and Niu (2011) identified the duration

between these measurements varied significantly. They found that there was a positive correlation between the statistical significance of the development of critical thinking ability of test participants and time intervals between testing and retesting.

The variability in testing regimes may be one factor used to explain why changes in critical thinking as measured by these tests seem to provide inconsistent results; some studies report significant changes while others report no significant changes in critical thinking (Behar-Horenstein and Niu, 2011). For example Carter *et al.* (2015) found studies which used the CCTST or the WGCTA did not all support the hypothesis of improved critical thinking with time in their longitudinal studies, with some studies reporting increases, and some studies reporting decreases or no change over time. Similarly, Huber measured the effect sizes of 124 longitudinal and cross sectional studies to find that effect sizes for the WGCTA and the CCTST ranged from -.65 to 2.22. A larger effect size ($r > .5$) would suggest over time a participants' critical thinking score improves due to maturity or training. Conversely, a small or negative effect size ($r < .1$) would suggest maturity or training were of little benefit, or to the individuals' detriment with respect to achievement on a critical thinking test (Cohen, 1988). The range of effect sizes suggest a negative relationship between the development of critical thinking and years of tertiary study in some empirical research while others displayed a very strong positive relationship, leaving cause for concern regarding the application of these tests.

The breadth of these reviews highlights the importance of experimental design when evaluating critical thinking. A review of 27 studies conducted by McMillan (1987) found that only seven studies demonstrated the significant changes in critical thinking. He concluded that tests which were purposely designed by the researcher are a better measure of critical thinking, as they specifically address the desired critical thinking learning outcomes, as opposed to commercially available tools which attempt to measure critical thinking as broad and generalised construct. Furthermore, of the seven studies which reported statistically significant changes in critical thinking, only two of these studies supported their findings by administering the test to control groups in addition to the treatment group. Additionally, the reviewer advises that studies with control groups be viewed with caution.

The importance of control groups and duration of studies is echoed by other reviewers who propose three research designs: true experimental (TE), quasi-experimental (QE) and pre-experimental (PE) (Behar-Horenstein and Niu, 2011). TE involves two or more treatment groups and control groups where participants have been randomly assigned to groups, whereas QE typically uses at least one treatment group and one control group where the participants have not been randomly assigned (Dooley, 2001, p.165). PE is the least rigorous experimental design in which no control group is used and participants for the treatment group utilise an entire cohort of students rather than randomly assigning them to the treatment group. PE experimental design is the most vulnerable to internal reliability issues, however analysis by Behar-Horenstein and Niu (2011) showed that 60% of studies take a PE approach, while 33% use a QE method and only 7% use TE method. They highlight that as the research design becomes more rigorous the proportion of statistically significant findings declined.

1.6 Considerations When Evaluating Critical Thinking Skills

As described above, there are certain key considerations regarding the development and usefulness of a critical thinking test. First of all, various sources of literature agree that the measurement of critical thinking needs to align closely with what the intervention(s) seeks to change (Abrami *et al.*, 2008; Carter *et al.*, 2015; Ennis, 1993; O'Hare and McGuinness, 2015). For example, of the empirical studies which used the WGCTA(-S) or the CCTST, very few offered a rationale for why they chose the test (Behar-Horenstein and Niu, 2011). Ennis (1993) proposes that an operational definition of critical thinking must be determined so that the assessment actually targets the investigator's view of critical thinking.

Ennis (1993) stated that assessment of critical thinking can serve several purposes. These purposes include assessing the level of students' critical thinking, providing students with feedback regarding their critical thinking, motivating students to develop their critical thinking, and informing educators of the effects their interventions had on students' critical thinking. Ennis goes on to say that no critical thinking assessment can cater to all these purposes at once and that the investigator must decide on the purpose of their assessment, whether it be a commercially available test or a non-standardised assessment.

There is a need to be aware of experimental design constraints. Such constraints, including the number of students to whom the test will be administered, must be considered for reasons such as cost and time required for evaluation (Ennis, 1993). Other considerations include how the research will be conducted. For example, if it will be possible to obtain a randomly assigned control group undergoing different treatments, or the duration of time between measurements (Behar-Horenstein and Niu, 2011). Finally there is evidence to suggest that students perform better on critical thinking assessments which are related to their discipline as they attach greater significance to the assessments (Garratt *et al.*, 2000; Halpern, 1998).

1.7 Research Question

The term critical thinking or expressions referring to critical thinking skills and behaviours such as 'analyse and interpret data meaningfully' can be found listed in the graduate attributes of many universities around world (Australian National University, 2015; Monash University, 2015; Ontario University, 2017; University of Adelaide, 2015; University of Edinburgh, 2017; University of Melbourne, 2015) . In Australia the inclusion of such graduate attributes in science is supported by many national initiatives such as the implementation of threshold learning outcome (TLO) 3.1 described in the Australian Learning and Teaching Academic Standards Statement, in order to develop skills that could be described by the term critical thinking. TLO 3.1 states that 'upon completion of a bachelor degree of science, graduates will critically analyse and solve scientific problems by gathering, synthesising and critically evaluating information from a range of sources' (Jones, Yates and Kelder, 2011).

Many studies highlight motivations for the development of higher education graduates' critical thinking skills from the perspective of students and employers. A study conducted by the Office of the Chief Scientist of Australia demonstrated employers' interest in critical thinking as a desirable graduate attribute (Prinsley and Baranyai, 2015). When they surveyed 1,065 employers representing a range of industries they found that employers considered critical thinking to be the second most important skill or attribute behind active learning (i.e. learning on the job). They also found that over 80% of respondents indicated critical thinking as

'important' or 'very important' as a skill or attribute in the workplace. The findings made by Prinsley and Baranyai (2015) are also observed internationally. A study of 263 employers from a variety of industries in the US recognised that leading innovation and change requires an aptitude for critical thinking (Desai *et al.*, 2016). In a survey of 400 US employers, 92% of respondents rated critical thinking as 'important' or 'very important' in an undergraduate degree, and the fifth most applied skill in the workplace (Jackson, 2010). These findings are indicative of the persistent needs of the job market for new graduates, and the emerging expectations that graduates are clearly able to demonstrate skills such as critical thinking (Lowden, Hall, Elliot and Lewin, 2011).

Published literature and government-supported studies both highlight the need for students to develop and articulate critical thinking skills in the competitive job market. In 2015, Graduate Careers Australia found that of the 45% of chemistry graduates available for full-time or part-time employment, only 66% of them had obtained employment in a chemistry-related field (Graduate Careers Australia, 2015). The remaining 34% of chemistry graduates obtained employment in fields unrelated to chemistry, such as account or finance. In these non-chemistry roles graduates are typically employed for their transferable or generic skills, such as critical thinking, rather than their chemistry discipline knowledge. A survey of 167 recent science graduates compared the development of a variety of skills developed whilst studying to the skills used in the work place (Sarkar, Overton, Thompson and Rayner, 2016). It found that 30% of graduates in full-time positions identified critical thinking as one of the top five skills they would like to have developed further within their undergraduate studies. Another study of 315 students from biological sciences, chemistry, and environmental management revealed an increase in the importance undergraduates placed on critical thinking skills as they progressed through their studies (Leggett, Kinnear, Boyce and Bennett, 2004). This cross-sectional study demonstrated that 55% of third year students rated critical thinking as highly important compared to 40% of first year students.

Students, governments and employers all recognise that not only is developing students' critical thinking an intrinsic good, but that it better prepares them to meet and exceed

employer expectations when making decisions, solving problems and reflecting on their own performance in graduate employment (Lindsay, 2015). Hence, it has become somewhat of an expectation from governments, employers and students that it is the responsibility of higher education providers to develop students' critical thinking skills. Yet, despite the clear need to develop students' critical thinking, measuring student attainment of critical thinking skills is inherently challenging, and rarely done in a meaningful way. There are ongoing debates regarding best practice for the development and evaluation of student critical thinking, as well as how students, academics and employers define critical thinking.

The growing need for developing the critical thinking skills of students at higher education institutions has led to the design and implementation of a breadth of teaching interventions and the development of a range of methods of assessing the impact of these interventions. Many of these assessment methods utilise validated, commercially available tests. However, there is evidence to suggest that if these assessments are to be used with tertiary chemistry students, the context of the assessments needs to be in the field of chemistry so that the students will attach significance to the assessment. Consequently, the students' performance on a critical thinking test will better reflect their actual critical thinking abilities.

Thus, an opportunity has been identified to develop a chemistry critical thinking test. Such a test could be broadly used to assist chemistry educators and chemistry education researchers in evaluating the effectiveness of teaching interventions designed to develop the critical thinking skills of chemistry undergraduate students. This leads to the core research question this thesis will address:

- How can a valid and reliable test be designed to measure undergraduate chemistry students' critical thinking skills independent of extensive chemistry knowledge, whilst set within a broad chemistry context?

To achieve this aim, an understanding of the functional definition of critical thinking within chemistry and identification of opportunities to develop critical thinking within the study of chemistry are required, as suggested by Ennis (1993). This leads to subsequent research questions:

- How do chemistry students, chemistry teaching staff and employers of chemistry graduates define critical thinking? What are the similarities and differences between these groups and how do their definitions compare to literature definitions?
- Where do chemistry students and chemistry teaching staff believe critical thinking is developed while studying chemistry at university and how do these views compare?

This thesis describes the research conducted to understand the definition(s) of critical thinking within the chemistry community, now published in the journal *Chemistry Education Research and Practice*. The thesis subsequently describes the development, reliability and validity studies to produce an instrument with which to measure undergraduate chemistry students' critical thinking skills.

Chapter 2 Methodology

This chapter will outline the methods for the development and evaluation of a chemistry critical thinking test aimed at undergraduate chemistry students with any level of experience, the Danczak-Overton-Thompson Chemistry Critical Thinking Test (DOT-CCTT). Section 2.1 and 2.2 present the theoretical framework and test theories which guided this research. As commercially available critical thinking tests are considered to evaluate a psychometric construct (Nunnally and Bernstein, 1994) there must be supporting evidence of their reliability and validity (DeVellis, 2012; Kline, 2005). An overview of the various forms of reliability and validity used in developing and evaluating these critical thinking tests will then be presented in Sections 2.3 and 2.4. This overview will be followed by a discussion of the methods of statistical analyses used and their rationale for inclusion in the study. A discussion of the decision to view statistical data non-parametric or 'not normal' data and the inherent assumptions and limitation will then follow. Finally, this chapter will describe the five stages of development of the DOT-CCTT, including a qualitative study published in the journal *Chemistry Education Research and Practice* in Sections 2.8 and 2.9, followed by detailed methodologies and descriptions of participants at each stage of the study.

2.1 Qualitative Research Theoretical Framework

There are two major parts of this thesis which require the consideration of theoretical frameworks. The first is the qualitative study of student, teaching staff and employers' perceptions of critical thinking. The second is with respect to how research participants engage with various iterations of the DOT-CCTT.

The first qualitative aspect of this research sought to understand two perceptions of critical thinking; (1) how students, teaching staff and employers define critical thinking, and (2), where did students and teaching staff believe undergraduate chemistry students have the opportunity to develop critical thinking skills. The second qualitative element of this thesis sought to examine how participants engaged with various iterations of the DOT-CCTT. It was interested in understanding the effect of using scientific terminology on a critical thinking test,

what information participants perceived as important, what they believed the questions were asking them, and to understand the reasoning underpinning their responses to questions on the DOT-CCTT. These questions aimed to understand individuals' perceptions of critical thinking and synthesise them into a generalizable truth. To this end, the core research theoretical framework was underpinned by constructivism (Ferguson, 2007). Constructivism is commonly employed in education research in cases where researchers are interested in understanding how participants make sense of an object or construct. For example: how do undergraduate chemistry students interpret gas laws or organic chemistry arrow mechanistic formalisms? Research questions that are investigated using constructivism are based on the theory that 'knowledge is constructed in the mind of the learner' (Ferguson, 2007, p. 28). This knowledge is refined and shaped by the learners' surroundings and social interactions in what is referred to as social constructivism (Ferguson, 2007, p. 29). As participant responses could be viewed through the lens that their perceptions of critical thinking are dependent on the context of the individuals' prior experience, social constructivism was suitable as a research theoretical framework for the qualitative studies within this thesis.

2.2 Test Theory Framework

With respect to the statistical analysis of the DOT-CCTT, it was important to distinguish which theoretical lens test scores were viewed through. The theories which dominate the literature are classical test theory and modern test theory, commonly referred to as item response theory (Kline, 2005). The score an individual obtains on a test is what is referred to as the raw score. That is, if the person were administered the test in an infinite number of testing sessions the raw score is the score they would on average obtain. The raw score is made up of the true score, the degree to which the individual actually possesses the trait or construct being tested, and the random error associated with the test (DeVellis, 2012). Classical test theory and item response theory delineate in relation to the respective theories' treatment of the raw score, the appraisal of question(s) difficulty, and the error associated with the measurement.

Classical test theory views a test as a whole, and therefore the test score is representative of a construct as a whole. Classical test theory assumes all questions are parallel, meaning the questions are independent of one another, contribute to the construct of interest equally, and the error associated with each question is the same (Kline, 2005). The reliability and validity of a test using classical test theory typically improves with a greater number of questions in the test. Difficulty analysis typically isn't conducted at question level but crude measurements can be conducted (Kline, 2005). Difficulty level analysis can be performed for dichotomous variable (questions with only two options) and is determined by the percentage of participants who obtain the correct response over the total number of participants. If the value is either 100% or 0% then the question is too easy or too challenging and does not provide any evidence to discriminate between participants. A difficulty analysis of 50% for example means that half of the participants would obtain the correct response, and the other half would obtain the incorrect response.

Conversely, item response theory does not view the test score as a whole, instead focusing on the contribution each question has on the construct of interest (Kline, 2005). Item response theory does not assume that all questions are parallel, and that the questions may be dependent on one another (Kline, 2005). Item response theory does not assume all test questions contribute equally to the construct of interest, or that the error associated with each question is equal. Item response theory does not require participants to undertake a large numbers of questions for purposes of reliability. Instead, reliability can be improved by altering and/or removing questions (Kline, 2005). The item response theory approach is therefore well suited to short tests (as short as five questions) administered to large numbers of participants (over 400 individuals). Most notably, item response theory considers participants ability for each question. This is achieved by plotting ability with up to three parameters which include item difficulty, how well a question may discriminate between certain groups, and accounting of the participants guessing (Kline, 2005). To obtain this level of question scrutiny, the aforementioned parameters need to adhere to characteristic curves and it is assumed that questions are locally independent of one another, meaning answering one part of a question

incorrectly would not inhibit a participant's ability to answer subsequent questions correctly (Kline, 2005).

Considering the need to restrict the number of questions on the DOT-CCTT (See Section 2.10), an item response theory approach was more appropriate. However, elements of classical test theory featured as part of the analysis. For example, item difficulty was conducted in the somewhat simplistic fashion of comparing the number of participants who obtained the correct response to a question over the total number of participants (See Section 2.11.3), and median scores were used to compare the performance of participants from different groups, for example comparing first year undergraduates and third year undergraduates (See Section 2.13.6). Furthermore, performance of different groups was investigated question by question to determine any emergent trends (Section 2.13.6). The item response theory analysis using more complex mathematics to account for parameters such as question difficulty, question discrimination, and the role of guessing fell outside of the scope of this study, however there are plans to apply these methodologies in further studies.

2.3 Reliability

In terms of psychometric testing, reliability is most often determined from test-retest reliability, internal reliability and inter-rater reliability (Nunnally and Bernstein, 1994). Test-retest reliability suggests that if a participant achieves a certain score on a test measuring a particular construct, assuming no change has occurred in the participant relative to the construct of interest, they will obtain a similar score upon repeating the test (Kline, 2005, pp. 168-171).

Internal reliability indicates whether questions and/or sections within a test are measuring the same scale or construct. The most commonly used method of evaluating internal reliability is to determine the internal consistency by calculating Cronbach's α . Cronbach's α is a single value which is determined from averaging the correlations of each question within a test against all other questions within the same test (Cronbach, 1951). Typically, a test is considered to have good internal consistency if $\alpha > .7$ (DeVellis, 2012, p. 109). If a test is made up of several sub-scales it is also recommended that the internal

consistency be determined from averaging the correlations of each sub-scale score within the test against all other sub-scale scores within the same test. It is also recommended that the internal consistency of each sub-scale be determined from averaging the correlations of each question within a sub-scale against all other questions within that same sub-scale. When determining internal consistency of sub-scale a value of $\alpha > .7$ is desirable, however, often when sub-scales of a test consist of ten or fewer questions a value of $\alpha > .7$ may not be achieved due to insufficient data points (Pallant, 2016, p. 104).

If the internal consistency of a test is determined to be less than $\alpha > .7$ and the test consists of sub-scales with ten or fewer questions each, it is advisable to determine corrected item-total correlations (CITC) (Pallant, 2016, p. 104). CITC provides a correlation of each question relative to the score of the sub-scale. If the CITC of a question is less than .3 it suggests that the question may not be contributing to the score of the sub-scale. Furthermore, it is advisable to determine the internal consistency of a section when a single question is removed from that scale or sub-scale. The internal consistency when a single question is deleted from a scale or sub-scale should be determined for all questions within that scale or sub-scale, and is often referred to as α if deleted. If the internal consistency value when the question is deleted from the calculation improves, it suggests that the question may not be contributing to the score of the scale or sub-scale, and may need to be removed from any further analysis (Kline, 2005, p. 176; Pallant, 2016, p. 104). Another common approach to measuring internal reliability is the split-halves method where a test is split in two halves, typically by dividing the test into even and odd questions, and effectively treating them as two tests (DeVellis, 2012, pp. 45-49). The internal consistency for both halves is calculated then used to determine the Spearman-Brown coefficient (ρ_{cc}) (Kline, 2005, p. 173). A test is said to exhibit good internal reliability when ρ_{cc} is greater than or equal to .8.

Finally, inter-rater reliability indicates how reliable a test is when administered and scored by different moderators. Tests which are comprised entirely of multiple choice questions typically do not require this form of reliability. However, inter-rater reliability of tests with open-ended responses do require this evaluation. To determine inter-rater reliability,

each test must be scored by at least two moderators and those scores are then used to determine Cohen's κ (Hallgren, 2012). Cohen's κ is a measure of agreement between assessors ranging from -1.00 to 1.00, where a value 1.00 is perfect agreement between assessors and a value of -1.00 is perfect disagreement.

2.3.1 Internal Reliability

Internal reliability throughout this research was measured from two perspectives. The first was to view the 30 questions of the DOT-CCTT as a single scale measuring the construct of critical thinking skill. The second was to view the DOT-CCTT as five separate constructs which collectively measure critical thinking skill. As suggested by (Pallant, 2016) when considering a test to be made up of several sub-scales or section, the internal reliability of the sub-scales (for example 'Making Assumptions') which make up a test should also be explored. Therefore, the internal consistency and the corrected item-total correlations (CITC) were determined for the sub-scales of DOT-CCTT, and the effect of deleting questions and/or sections of the test has on the internal consistency on the DOT-CCTT. To further explore the internal reliability of the DOT-CCTT a split halves analysis was also conducted dividing the test into even and odd questions. The correlation between the two halves was determined using the Spearman-Brown Coefficient.

2.3.2 Item Difficulty Analysis

The number of students who answered each question on the DOT-CCTT correctly was used as a visual indication of the difficulty of each question and used to perform simple item difficulty analyses as described by Loo and Thorpe (1999). The approach described by the authors offered a way to appraise the difficulty of the questions of the DOT-CCTT, whereby the most desirable question difficulty, as measured by the percentage of participants answering the question correctly, lay half way between the probability of answering the question correctly by chance and 100%. For example, the format of the 'Making Assumptions', 'Analysing Arguments', 'Testing Hypotheses', and 'Drawing Conclusions' sections of the DOT-CCTT (see Section 2.10) were multiple choice questions with two possible answers.

Therefore, there was a 50% chance a participant would select the correct response by guessing and so the desirability for these questions was 75%. The authors also suggest that questions where more than 90% of participants obtain the correct response suggests the question is not difficult enough to discriminate between participants. Likewise, if the percentage of participants who respond to a question correctly is lower than the probability of obtaining the correct response by chance (for example less than 50%), the question is poorly constructed or too difficult to discriminate between participants. In the case of the 'Developing Hypotheses' section of the DOT-CCTT (see Section 2.10), 66% of participants obtaining the correct response would reflect the most desirable difficulty, and if less than 33% of participants obtain the correct response it would be an indication the question requires revision.

The median score of participants who answered a given question correctly and the median score of participants who answered the same question incorrectly were used in Mann-Whitney *U* tests to determine if answering a given question correctly was a significant predictor of obtaining a higher score on the DOT-CCTT. The strength of the relationship between responding to questions correctly or incorrectly and the participant obtaining a high score on the DOT-CCTT was then determined by calculating the effect size (*r*) of each question from the data obtained from the Mann-Whitney *U* tests. A larger effect size ($r > .5$) in this instance would suggest that a participant who responds correctly to the given question is more likely to obtain a higher score on the DOT-CCTT. Conversely, a small effect size ($r < .1$) would suggest that obtaining the correct response on the given question would not indicate the participants who would obtain a high score on the DOT-CCTT. For example, if a question was found to have an effect size of .53 when comparing the median DOT-CCTT score of participants who answered that question correctly to the median DOT-CCTT score of participants who answered that question incorrectly in a Mann-Whitney *U* test, answering the question correctly would be a strong predictor of a participant being part of the group who obtained a high median score on the DOT-CCTT.

2.3.3 Test-Retest Reliability

Test-retest reliability is typically achieved by performing a paired *t*-test or Wilcoxon signed rank test (Pallant, 2016, pp. 234-236, pp. 249-253), which determines whether there is any significant difference between the test scores taken at different time intervals with respect to the same participants. When the scores of the tests taken at different times have no significant difference, as determined by a *p* value greater than .05, the test can be considered to have acceptable test-retest reliability (Pallant, 2016, p. 235). It is important to clarify that acceptable test-retest reliability does not imply that test attempts are equivalent. Rather, good test-retest reliability suggests the precision of test to measure the construct of interest is acceptable.

2.4 Validity

Evaluation of the validity of psychometric tests typically consists of content validity, criterion validity and construct validity. Construct validity can be broken down further into convergent and discriminate validity. Content validity is used to determine whether the right questions are being asked and if there are enough questions correctly measuring the intended construct (Nunnally and Bernstein, 1994). This is sometimes referred to as face validity (Stein *et al.*, 2007). Content validity is reflected in Ennis' (1993), Dressel's and Mayhew's (1954) notion that a test should measure the outcome(s) of intended teaching interventions. For example, content validity can ask: does a mid-semester test include questions from all the previous weeks? Content validity is most commonly assessed through focus groups and interviews (Kline, 2005, pp. 202-203). Typically a group of experts including education, psychology and discipline specific academics will attempt the test, followed by interviews or focus groups. From the qualitative analysis of this data, tests are edited in order to improve content and construct validity before a larger scale study is conducted with the intended participants. The participants of this larger study are typically interviewed to gain an understanding of the cognitive processes they employed and any feedback they may have regarding the test.

Criterion validity asks whether the test is good predictor of the participants' behaviour relating to the construct intended to be measured (Kline, 2005, p. 203). This form of validity is seldom explicitly discussed in the literature, but it is often alluded to with the assumption that there will be a positive correlation between test scores and years of tertiary education or experience (Daly, 2001; Gadzella *et al.*, 2006; Giancarlo and Facione, 2001). Criterion validity typically takes the form of longitudinal studies of the same participants over time, or cross-sectional studies of several cohorts at one specific point in time usually differentiated by a key variable, for example years of tertiary study (Gadzella *et al.*, 2006; Hassan and Madhum, 2007; Loo and Thorpe, 1999). In both of these instances tests scores are compared using a *t*-test or Mann-Whitney *U* test. Good validity in these instances is shown by a $p < .05$ which indicated that there is a statistically significant difference between the median scores of the two groups being considered (for example a first year group and an academic group). Furthermore, effect sizes greater than $r = .5$ suggests a large effect size (Cohen, 1988), meaning there is strong association between the variables (for example score on a critical thinking test and years of tertiary study).

Construct validity simply asks whether the test is measuring the intended construct. In psychometric testing this is typically determined through convergent validity and discriminate validity. Convergent validity is how well the test been developed compares with a test which also measures the intended construct and has proven reliability and validity (Nunnally and Bernstein, 1994). Typically, participants will complete the test being developed alongside a test which has already undergone validity and reliability testing (Tiruneh *et al.*, 2016; Watson and Glaser, 2006). The scores of these tests are compared using Pearson's product of moment correlation or a Spearman's ρ correlation and significant ($p < .05$) values of correlation suggest a strong correlation between tests (Cohen, 1988).

Discriminate validity asks whether the test is only measuring the intended construct or whether it is dependent on other variables (Nunnally and Bernstein, 1994). Common examples in the literature include the assessment of the dependence of university entrance scores or years of age relative to scores on the critical thinking test of interest (Macpherson and Owen,

2010; O'Hare and McGuinness, 2015). The correlation between the variable of interest and the test scores can be determined statistically by using Pearson's product of moment correlation or a Spearman's ρ correlation, where the correlation value can range anywhere from -1.00 to 1.00, indicating whether the variable and the score are inversely or directly correlated. For example, if the relationship between sex and score on a critical thinking test was found to have a Spearman's ρ value of 0.86, the variable of sex would be considered a variable which contributes to the score on the critical thinking test and that performance on the test is dependent on the participants' sex. This would be considered poor discriminate validity (DeVellis, 2012, p. 69; Kline, 2005, p. 286).

2.4.1 Convergent Validity

Convergent validity was determined using a reliable and valid commercially available test comparable to the DOT-CCTT in length, style of question, and the areas of critical thinking that both tests intend or claim to measure (See Section 2.12.3 for further discussion). A Spearman's ρ correlation was used to compare the participants' median scores obtained on the DOT-CCTT with their median scores obtained on the commercially available test. The ρ value can range from +1.00 to -1.00. A positive value of ρ would indicate that an individual who performed well on the commercially available test should also perform well on the DOT-CCTT. As the value of ρ approaches positive the correlation between performances on the tests becomes stronger (Cohen, 1988). Furthermore, if the value of ρ is considered significant ($p < .05$) further weight is added to support an argument of good convergent validity of the DOT-CCTT. Conversely, if ρ is a negative value, it would indicate that a strong performance on the commercially available test would lead to poor performance on the DOT-CCTT, suggesting poor convergent validity of the DOT-CCTT.

2.4.2 Criterion Validity

In order to determine the criterion validity several Mann-Whitney U tests were conducted to establish whether there were any differences between groups and the median scores obtained on the DOT-CCTT. For example, comparing the median scores of first year

students and third year students. It was predicted that if the DOT-CCTT had good criterion validity then the Mann-Whitney *U* test would detect statistically significant differences in DOT-CCTT scores between groups. Retrospectively, it is clear that a Kruskal-Wallis test would have been a more efficient approach to determining criterion validity, as it is a single test which can measure a continuous variable, such as score on the DOT-CCTT, for three or more groups (Pallant, 2016, p. 236). At the time of analysis Mann-Whitney *U* tests were used for criterion validity rather than a single Kruskal-Wallis test due to familiarity with the technique. In hind sight, the approach using Mann-Whitney *U* tests was more involved but nonetheless equally valid as a Kruskal-Wallis Test. As stated by Pallant (2016, p. 236) “It [Kruskal-Wallis test] is similar in nature to the Mann-Whitney *U* test...but it allows you to compare more than just two groups” at a time.

2.4.3 Discriminate Validity

In order to determine the discriminate validity of the DOT-CCTT, demographic data was collected regarding the participants’ age, their previous academic achievement as measured by the participants tertiary entrance scores (within Australian as an Australian Tertiary Admission Rank [ATAR] score), and which university the participants attended in the case of the third year undergraduate students. The relationship between score obtained on the DOT-CCTT and age was investigated using Spearman’s Rank-order correlation coefficient. To determine if the DOT-CCTT scores were independent of previous academic achievement the relationship between score obtained on the DOT-CCTT and ATAR score was investigated using Spearman’s Rank order correlation coefficient. The validity of the DOT-CCTT outside of Monash University was determined by comparing the median scores on the DOT-CCTT of third year students from Monash University with the median scores on the DOT-CCTT of students from Curtin University using a Mann-Whitney *U* Test.

2.4.4 Content Validity

The data collected from focus groups was analysed qualitatively through a social constructivist theoretical frameworks, as discussed in Section 2.1. The discussions were

recorded with permission of the participants and transcribed verbatim into Microsoft Word. The transcripts were then imported into NVivo version 11 and an initial analysis was performed to identify emergent themes. The data then underwent a second analysis to ensure any underlying themes were identified. Then via a third review of the data and using a redundancy approach, similar themes were combined resulting in distinct themes which were used for subsequent coding of the transcripts (Bryman, 2008).

2.5 Use of Non-Parametric Tests versus Parametric Tests

The literature suggests that in order to treat test scores as normal data one of the requirements is that dependent variables use a continuous scale (Pallant, 2016, p. 207). The questions which made up the DOT-CCTT score were either dichotomous or multiple choice question, as will be discussed in Section 2.10. Due to their format the responses to each question were considered categorical variables. Therefore the DOT-CCTT data was treated as not-normal or non-parametric, which restricted the reliability and validity statistical analyses that could be conducted on the data. For example, when comparing the score of groups such as first year students and third year students, treatment of the data as non-parametric required the comparison of median scores using Mann-Whitney *U* tests rather than the more sensitive comparison of mean scores using *t*-tests. Furthermore, when sample cohorts are small, such as the cross-sectional analysis of DOT-CCTTv2 in Section 2.12 where $n = 20$ it is also recommended that the data be treated as non-parametric (Pallant, 2016).

2.5.1 Assumptions Regarding Type 1 and Type 2 Errors

When discussing the statistical analyses within this thesis, it is important to acknowledge the limitations related to Type 1 and Type 2 errors when determining statistical relationships between variables (Pallant, 2016, p. 209) such as DOT-CCTT scores, years of tertiary education, age, previous academic achievement, sex or institution the participant attended. When the degree of overlap between variables (such as score on the DOT-CCTT and years of education), as measured by a *p* value in tests such as the Mann-Whitney *U* test, the Wilcoxon signed rank test or Spearman's rank-order correlation coefficient, is greater

than .05, any difference between variables is considered to be due to chance which is referred to as the null hypothesis (Pallant, 2016, p. 209). The null hypothesis is retained in statistical analyses unless the associated p value, is found to be less than .05. When p is less than .05 any relationship between variables is considered to be not due to chance and is therefore statistically significantly different, in which case the null hypothesis is not retained.

The Type 1 error is a rejection of the null hypothesis, when in fact a significant difference between variables does not truly exist and the null hypothesis should be retained. Conversely, the Type 2 error is the retention of the null hypothesis and a failure to identify a significant difference between variables which actually exists. Type 1 and Type 2 errors are inversely related such that reducing the likelihood of committing the Type 1 error increases the likelihood of committing the Type 2 error (Pallant, 2016, p. 209), and *vice versa*.

Statistical analyses conducted as part of this research were unlikely to have made the Type 1 error as the data collected throughout these studies was treated as non-parametric data and more likely prone to the Type 2 error. As described above, all data was treated as non-parametric due to the fact that the questions of the DOT-CCTT only had two or three response options each (Pallant, 2016, p. 104). Since non-parametric analyses such as the Mann-Whitney U test were used throughout the study of the DOT-CCTT, as opposed to parametric equivalents, such as the t -test which is more sensitive in detecting statistical significance compared to the Mann-Whitney U test, it is more likely that the null hypothesis and relationships between variables were underestimated rather than overestimated. The reduced sensitivity inherent in treating data as non-parametric data has been considered acceptable for this research as it would be more academically responsible to maintain conservative claims regarding the sensitivity of the DOT-CCTT, rather than overstate the capabilities of the test.

2.6 Ethics

All participants throughout the studies described herein were informed that their participation was voluntary, anonymous, would in no way affect their academic or professional

records, and that they were free to withdraw from the study at any time. Participants were provided with an explanatory statements outlining these terms and all procedures were in accordance with Monash University Human Research Ethics Committee (MUHREC) regulations (project numbers CF15/560 – 2015000258 and CF16/568 – 2016000279) (See Appendix A and B).

2.7 Methodology Overview

The development and validation of the DOT-CCTT was performed in five stages highlighted in Figure 2. 1. This section provides an introduction to the five stages of development and evaluation. Each stage of development and evaluation will be discussed in detail in subsequent sections of this chapter.

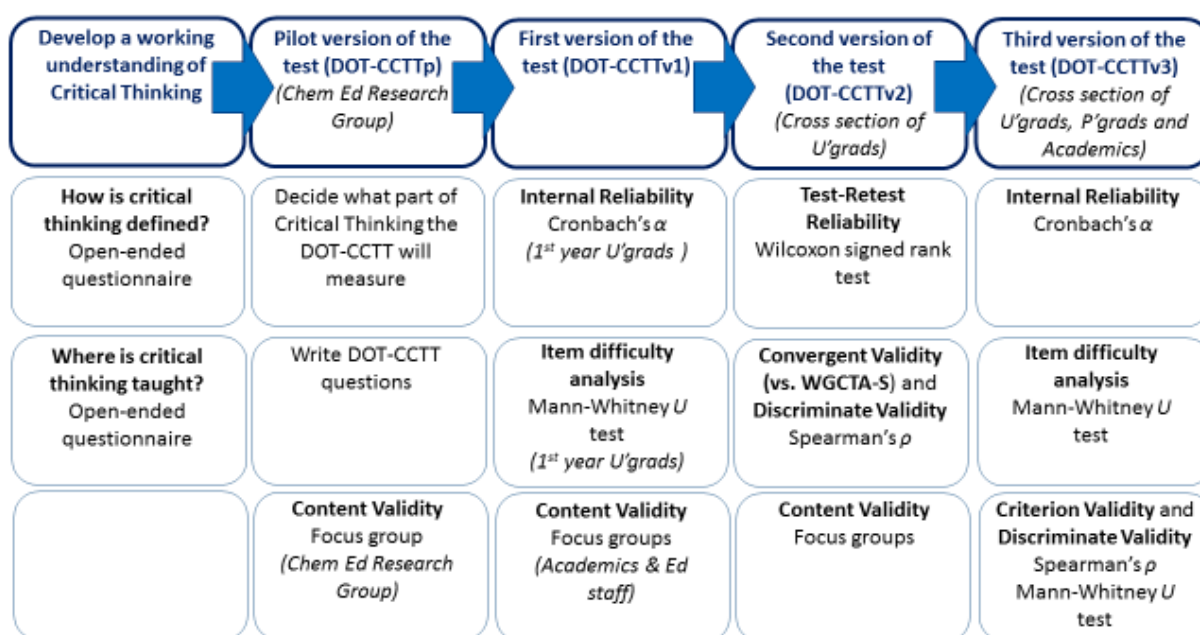


Figure 2. 1 Flow chart of methodology, consisting of writing and evaluating the reliability and validity of iterations of the DOT-CCTT

The first stage of test development was to gain an understanding of how critical thinking is defined and where it is taught in the context at Monash University. These findings were then compared with definitions and critical thinking pedagogies described in the literature. Comparing the data obtained from the Monash University context alongside the literature facilitated the development a functional definition of critical thinking which informed the

development of the DOT-CCTT throughout this research. The second stage of development used this data to decide which elements of critical thinking the DOT-CCTT would test and, therefore, which commercially available test(s) were suitable as guides for question development.

Stage 3 focused on determining the internal reliability and item difficulty through administering the first version of the DOT-CCTT (DOT-CCTTv1) to first year undergraduate chemistry students at Monash University. Concurrently, content validity of the DOT-CCTTv1 was assessed with the assistance of the Monash Science and Technology Education Researchers community of practice. Stage 4 of the development and evaluation of the DOT-CCTT was conducted with a cross-sectional group of undergraduate chemistry students from Monash University, ranging from first year to fourth year students. Test-retest reliability, convergent validity and content validity of the DOT-CCTTv2 was studied via this group.

The fifth stage of the research focused on determining the internal reliability, item difficulty, criterion validity and discriminate validity of the DOT-CCTTv3. The test was administered to first year and third chemistry undergraduates at Monash University, third year chemistry undergraduates at Curtin University, and a range of PhD students, post-doctoral researchers and academics drawn from Monash University and an online chemistry education research community of practice.

2.8 A Qualitative Analysis of Students', Teaching Staff and Employers'

Views of Critical Thinking

When designing interventions aimed to develop and assess critical thinking it is valuable to appreciate the definitions participants hold when describing critical thinking (Ennis, 1993). Furthermore, an understanding of the contexts in which participants think they develop critical thinking helps guide current and future teaching practices as described by the theory of constructive alignment (Biggs, 2012). Constructive alignment suggests that learning outcomes, teaching activities and learning assessment align in a way that students can comprehend what they are learning, how they are going to learn it, and that the assessment

will allow the students to demonstrate the learning outcomes. The research described in the following paper aimed to understand students' conceptualisation of critical thinking and where in the curriculum they believed that those skills were developed. As teaching associates and academics provide the learning opportunities for students to develop critical thinking skills, this research also investigated their definitions of critical thinking and perceptions of where those skills are developed. In a study on critical thinking it is important to know where the main stakeholders are positioned in terms of their conceptualisation of critical thinking in the context of chemistry.

In order to achieve these goals an open-ended questionnaire was designed and administered to all three year levels of Monash University undergraduate chemistry students, teaching staff at Monash University, and teaching staff from an online community of higher education practitioners in the field of chemistry, in 2015. The questionnaire contained open-ended fields asking students and teaching staff to define critical thinking and to identify where critical thinking is developed while studying chemistry. To add breadth to the study, employers of science graduates were invited to provide their definition of critical thinking in 2016 via an online questionnaire. Employers were not asked to respond to the question regarding where students develop their critical thinking. The data obtained from these questionnaires was analysed through a constructivist research framework as described in Section 2.1.

Students and teaching staff were asked several additional questions which were not used in the paper (see Appendix C and D). These questions were not used as the results provided little insight. For example, 99% of participants either agreed or strongly agree that critical thinking was important to develop at university using a Likert scale question. Questionnaire data did not contain any identifying information, making follow-up interviews impossible. If a similar study were to be conducted in future, providing a field for contact information would assist in conducting interviews to probe the qualitative data of these questionnaires further.

There were several similarities and differences between students and teaching staff, as described in the following journal article. The key finding of the study was that students,

teaching staff and employers all identified deductive logic elements of critical thinking, such as 'analysis' and 'problem solving'. However, inductive logic elements, such as 'judgement' or 'inference', typical of the literature regarding the definition of critical thinking (Facione, 1990; Halpern, 1996b), were seldom described by students. This study suggests that students possess a narrow view of critical thinking based almost solely on problem solving. This is may be due to the students' educational context in which they have had the opportunity to experience critical thinking thus far. However, the development of judgement and inference skills are required by the competent chemistry graduate as highlighted by employers and universities around the world (Australian National University, 2015; Chapman and O'Neill, 2010; Desai *et al.*, 2016; Jones *et al.*, 2011; Monash University, 2015; Ontario University, 2017; Prinsley and Baranyai, 2015; University of Adelaide, 2015; University of Edinburgh, 2017; University of Melbourne, 2015). Therefore, a university education should expose students to develop the broader elements of critical thinking, even if these are not explicitly recognised by the students.

Students strongly identified the practical environment and inquiry-based activities as opportunities to develop critical thinking. Academics and teaching associates identified additional opportunities which included authentic research experiences, such as critiquing literature and conducting research projects. The teaching activities described by the teaching staff were more reflective of teaching at higher year levels, but students in their first year of chemistry recognised these learning opportunities when they were presented to them. Students may have difficulty recognising opportunities to develop their critical thinking skills outside of inquiry-based practical activities. Students may benefit from more overt approaches to the development of critical thinking, such as round table discussion of chemistry critical thinking problems in a tutorial setting (Garratt *et al.*, 2000) or a mixed approach where syllogistic reasoning is explicitly taught alongside the chemistry curriculum (Jacob, 2004). A detailed description of the methodology, analysis and discussion of this work can be found in the following paper published in 2017 in the peer-reviewed journal *Chemistry Education Research and Practice*.



Cite this: *Chem. Educ. Res. Pract.*,
2017, 18, 420

'What does the term Critical Thinking mean to you?' A qualitative analysis of chemistry undergraduate, teaching staff and employers' views of critical thinking

S. M. Danczak,* C. D. Thompson and T. L. Overton

Good critical thinking is important to the development of students and a valued skill in commercial markets and wider society. There has been much discussion regarding the definition of critical thinking and how it is best taught in higher education. This discussion has generally occurred between philosophers, cognitive psychologists and education researchers. This study examined the perceptions around critical thinking of 470 chemistry students from an Australian University, 106 chemistry teaching staff and 43 employers of chemistry graduates. An open-ended questionnaire was administered to these groups, qualitatively analysed and subsequently quantified. When asked to define critical thinking respondents identified themes such as 'analysis', 'critique', 'objectivity', 'problem solving', 'evaluate' and 'identification of opportunities and problems'. Student respondents described the smallest number of themes whereas employers described the largest number of themes. When asked where critical thinking was developed during the study of chemistry students overwhelmingly described practical environments and themes around inquiry-based learning. When teaching staff were asked this question they commonly identified critiques, research, projects and practical environments to some extent. This research highlights that there is only limited shared understanding of the definition of critical thinking and where it is developed in the study of chemistry. The findings within this article would be of interest to higher education teaching practitioners of science and chemistry, those interested in development of graduate attributes and higher order cognitive skills (HOCS) and those interested in the student and employer perspectives.

Received 19th December 2016,
Accepted 13th February 2017

DOI: 10.1039/c6rp00249h

rsc.li/cerp

Introduction

The development of critical thinking is a long standing goal of education at all levels (primary, secondary and tertiary) and a virtue valued by wider society. The importance of critical thinking is commonly referred to in literature discussions regarding employability and transferable skills (Ghulam and David, 1999; Leggett *et al.*, 2004; Sarkar *et al.*, 2016). In Australia, Group of Eight Universities list critical thinking and reasoning among the attributes of their science graduates (Australian National University, 2015; Monash University, 2015; The University of Adelaide, 2015; The University of Melbourne, 2015). The need for continued improvement in graduate employee higher order thinking can be seen through national initiatives (for example in the USA, UK, Canada and Australia) developing minimum learning requirements referred to as threshold learning outcomes (Wilson *et al.*, 1997; Pithers and Soden, 2000; Tapper, 2004; Jones *et al.*, 2011).

As the need for innovation, and anticipating and leading change continues to grow, employers recognise the importance of critical thinking and critical reflection (Desai *et al.*, 2016). It has become an expectation that graduates are able to demonstrate a range of transferable skills such as critical thinking (Lowden *et al.*, 2011). In a survey of 400 US employers, 92% of respondents rated critical thinking as 'important' or 'very important' in an undergraduate degree and the fifth most applied skill in the work place (Jackson, 2010a).

A recent study commissioned by the Office of the Chief Scientist of Australia surveyed 1065 employers representing a range of industries (Prinsley and Baranyai, 2015). Over 80% of respondents indicated critical thinking as 'important' or 'very important' as a skill or attribute in the workplace. Critical thinking was considered the second most important skill or attribute behind active learning. In 2012 Graduate Careers Australia found that of the 45% of chemistry graduates available for full-time or part-time employment, only 66% had obtained employment in a chemistry related field (Graduate Careers Australia, 2015). These findings suggest that skills which may be transferable to a range of employment

School of Chemistry, Monash University, Victoria 3800, Australia.
E-mail: Stephen.danczak@monash.edu

settings, such as critical thinking, are worthwhile developing at the tertiary level.

The definition of critical thinking

The definition of critical thinking is frequently discussed in the literature, particularly among philosophers, psychologists and education researchers. From a philosophical perspective a comprehensive dialogue regarding critical thinking emerged in the form of the Delphi report (Facione, 1990). This report summarised a year-long discussion between 47 academics from philosophy, education, social sciences and physical sciences. They arrived at a general consensus that critical thinking is 'purposeful, self-regulatory judgement which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgement is based' (Facione, 1990, p. 2).

The report concluded that a person who exhibits good critical thinking is in possession of a series of cognitive skills and dispositions. The consensus of the Delphi experts was that a good critical thinker is proficient in the skills of interpretation, analysis, evaluation, inference, explanation and self-regulation (Facione, 1990). Furthermore, the report stated that a good critical thinker demonstrates a series of dispositions which is required for the individual to utilise the aforementioned skills. According to the report a 'good critical thinker, is habitually disposed to engage in, and to encourage others to engage in, critical judgement' (Facione, 1990, p. 12). These dispositions were later categorised into inquisitiveness, open-mindedness, systematicity, analyticity, truth seeking, critical thinking self-confidence and maturity (Facione, 1990).

Cognitive psychology and education research take a more evidence based approach to defining critical thinking and the skills and dispositions that it encompasses. The term critical thinking itself is often used to describe a set of cognitive skills, strategies or behaviours that increase the likelihood of a desired outcome (Halpern, 1996; Tiruneh *et al.*, 2014). Dressel and Mayhew (1954) suggested it is educationally useful to define critical thinking as the sum of specific behaviours which could be observed from student acts. These critical thinking abilities are identifying central issues, recognising underlying assumptions, evaluating evidence or authority and drawing warranted conclusions.

Psychologists typically explored and defined critical thinking *via* a series of reasoning schemas; conditional reasoning, statistical reasoning, methodological reasoning and verbal reasoning (Nisbett *et al.*, 1987; Lehman and Nisbett, 1990). Halpern (1993) refined the cognitive psychologists' definition of critical thinking as the thinking required to solve problems, formulate inferences, calculate likelihoods and make decisions. Halpern listed a series of skills and dispositions required for good critical thought. Those skills are verbal reasoning, argument analysis, thinking as hypothesis testing, understanding and applying likelihood, uncertainty and probability, decision making and problem solving (Halpern, 1998). The dispositions Halpern described are a willingness to engage and persist with complex tasks, habitually planning and resisting impulsive

actions, flexibility or open-mindedness, a willingness to self-correct and abandon non-productive strategies and an awareness of the social context for thoughts to become actions (Halpern, 1998). Glaser (1984) further elaborated on the awareness of context to suggest that critical thinking requires proficiency in metacognition.

In the case of science education there is often an emphasis of critical thinking as a skill set (Bailin, 2002). There are concerns that from a pedagogical perspective many of the skills or processes commonly ascribed as part of critical thinking are difficult to observe and therefore difficult to assess. Consequently, Bailin suggests that the concept of critical thinking should explicitly focus on adherence to criteria and standards to reflect 'good' critical thinking (Bailin, 2002, p. 368).

Recent literature has lent evidence to the notion that there are several useful definitions of critical thinking of equally valuable meaning (Moore, 2013). The findings of this work identified themes such as 'critical thinking: as judgement; as scepticism; as originality; as sensitive reading; or as rationality.' The emphasis with which these themes were used was dependent on the teaching practitioners' context.

Can critical thinking be taught?

How critical thinking is taught and the extent to which critical thinking skills may be transferable between disciplines is a highly contentious issue. Teaching critical thinking dates back to ancient Greece in the philosophies of Socrates and Plato (Mann, 1979). Studies in the early to mid-twentieth century suggested that students are only able to think critically within their specialised discipline (Thorndike and Woodworth, 1901a, 1901b, 1901c; Inhelder and Piaget, 1958; Wason, 1966) and, therefore, to teach critical thinking in an abstract environment provides no educational benefit to the student's capacity to think critically (McPeak, 1981).

In later years cognitive psychology lent evidence to the argument that critical thinking could be developed within a specific discipline and those reasoning skills were, at least to some degree, transferable to situations encountered in daily life (Lehman *et al.*, 1988; Lehman and Nisbett, 1990). This led to a more pragmatic view that the best critical thinking occurs within ones area of expertise, termed domain specificity (Ennis, 1990), however critical thinking can still be effectively developed with or without content specific knowledge (McMillan, 1987; Ennis, 1989). However the debate regarding the dependence of content specific knowledge in the development of critical thinking continues to be discussed (Moore, 2011; Davies, 2013).

Attempts to teach critical thinking are common in the chemistry education literature. These range from writing exercises (Oliver-Hoyo, 2003; Martineau and Boisvert, 2011; Stephenson and Sadler-Mcknight, 2016), inquiry-based projects (Gupta *et al.*, 2015), flipped lectures (Flynn, 2011) and open-ended practicals (Klein and Carney, 2014) to gamification (Henderson, 2010) and work integrated learning (WIL) (Edwards *et al.*, 2015). While this literature captures that critical thinking is being developed, it seldom discusses the perception of the students.

This study aimed to identify the perceptions of critical thinking of chemistry students, teaching staff and employers.

The study investigated how each of these groups define critical thinking and where students and teaching staff believed critical thinking was developed during the study of chemistry.

Method

The research aims were achieved *via* qualitative analysis of open-ended questionnaire data collected in either paper or digital formats.

Data collection instrument

An open ended questionnaire was designed and administered to all three year levels of Monash University undergraduate chemistry students in 2015. The questionnaire contained questions regarding demographic data (age and gender) and two open-ended fields asking ‘What does the term “Critical Thinking” mean to you?’ (Q1) and ‘Can you provide an example of when you have had the opportunity to develop your critical thinking while studying chemistry?’ (Q2a) All participants were informed that their participation was voluntary, anonymous and would in no way affect their academic records. Participants were provided with an explanatory statement outlining these terms and all procedures were in accordance with Monash University Human Research Ethics Committee (MUHREC) regulations (project number CF15/560 – 2015000258).

A similar questionnaire was administered in hard copy to the teaching associates (TAs) and academics within the School of Chemistry at Monash University and *via* an online format to a different cohort from a range of institutions. The questionnaire consisted of items asking participants to identify teaching activities undertaken within the previous year, and at which year levels they taught these activities. They were asked open-ended questions which aligned with the student questionnaire: ‘What does the term “Critical Thinking” mean to you?’ (Q1) and ‘Can you provide an example of when you have provided students with the opportunity to develop their critical thinking while studying chemistry?’ (Q2b).

Employers were contacted directly *via* email and provided with a link to an online questionnaire. The questionnaire consisted of four open-ended question: ‘What does the term “Critical Thinking” mean to you?’ (Q1) and three demographic questions regarding which country the participant’s organisation was based, which sector their business was in and the highest qualification the participant held.

Student participants

The first year cohort was drawn from two units: Chemistry I and Advanced Chemistry I consisted of approximately 1000 students. Chemistry I is a general chemistry course with a mixed cohort of science students (880 in 2015), some of whom completed chemistry studies in high school and some who studied science but have not previously studied chemistry. Advanced Chemistry I consisted only of students who have completed chemistry in high school. Advanced Chemistry I covers the same content as Chemistry I with additional time in laboratory sessions. All first

year participants were provided with the questionnaire at the conclusion of a compulsory laboratory safety induction session during orientation in the first week of semester one. This data was considered a representative random sample of first year chemistry students as the induction session was a prerequisite of all students commencing study in the first year chemistry laboratory.

The second year cohort consisted of 359 students from Synthetic Chemistry I, a course focused on organic and inorganic synthetic techniques from practical and theoretical perspectives. This course is a core unit for any student pursuing a chemistry major. Participants were provided with the questionnaire at the end of a practical session during the first two weeks of semester one. The practical activity conducted within this time was known to typically only take students three of the four hours allocated to the practical session. As the activity was a compulsory part of the course, and given it was an essential part of the chemistry major, this cohort could be considered a representative random sample of second year chemistry major students.

Finally, the third year cohort was drawn from 84 students studying Advanced Inorganic Chemistry. This course builds on the theoretical knowledge and practical skills developed in Synthetic Chemistry I, focusing specifically on inorganic chemistry. Typically students completing a chemistry major undertook this unit but alternative courses were available. Participants were provided with the questionnaire during practical sessions in the first four weeks of semester and encouraged to complete it during the session. Since the activities in these sessions were very demanding and time was generally scarce for these students the sampling was regarded as convenient. Furthermore as not all chemistry majors may have undertaken Advanced Inorganic Chemistry the data obtained from this cohort may be non-representative.

Teaching staff participants

47 TAs from Chemistry I and Advanced Chemistry I were provided with the questionnaire at the conclusion of a compulsory laboratory safety induction during orientation of semester one. As such the data obtained from this cohort could be considered a representative random sample of TAs who taught at a first year undergraduate level.

A senior TAs and academics cohort consisted of academic staff and TAs with several years teaching experience. These academics and senior TAs typically taught chemistry courses other than Chemistry I or Advanced Chemistry I. 12 individuals were approached during semester one of 2015 and were advised to return the questionnaire *via* unlabelled internal mail.

Finally an online academic cohort consisted of around 300 members of a chemistry education email discussion group predominately from the UK and Europe. These participants received a link to an online version of the questionnaire sent *via* a third party.

All TAs and academic staff were advised their participation was voluntary and they could opt out by not completing the questionnaire in accordance with MUHREC regulations. All senior TAs, academics and online academics were previously known to highly value the scholarship of teaching thus

increasing the likelihood of their participation. Consequently this would be considered a non-representative and convenient sample of experienced teaching staff.

Employer participants

Participants were drawn from a list of respondents who were known to have previously participated in similar qualitative research (Sarkar *et al.*, 2016). Over 200 employers on this list were contacted but the number of responses was quite low (21%). The data from this cohort was a convenient sample and non-representative of all employers of chemistry graduates. All participants were informed that their participation was voluntary and they could opt out by not completing the questionnaire in accordance with MUHREC regulations.

Research theoretical framework

With respect to the open-ended questions posed to the various groups a realist world view was adopted (Edmunds and Brown, 2013). The philosophical framework informing this work was that of constructivism (Matthews, 1993). Constructivism postulates that individuals construct their meaning of concepts such as critical thinking from their experiences and interactions with the world around them. (Lemanski and Overton, 2011). The underpinning assumption was that there may be many truths regarding critical thinking which may be gleaned from data collected in a qualitative manner.

The data was analysed qualitatively and the next stage involved quantification of that qualitative analysis. The qualitative data was analysed with no prior assumptions regarding the number of ways in which individuals may think about critical thinking. The qualitative analysis was then quantified to identify whether there were any common ways in which individuals experienced critical thinking. The nature of these commonalities was not assumed however a retrospective comparison with the literature informed the inferences drawn from the data.

Data analysis

The responses from 470 undergraduate, 40 first year TA, 12 senior TA and academic, 55 online academic and 43 employer questionnaires were transcribed verbatim. This data was then imported into the qualitative analysis tool Nvivo version 10.

The questionnaire data for each cohort was imported into Nvivo as seven separate 'sources': first year students (A), second year students (B), third year students (C), TAs (D), senior TAs and academics (E), online academics (F) and employers (G). These cohorts were then merged into three major groups. Students, consisting of A, B and C, teaching staff consisting of D, E and F and employers (G).

Six chemistry education researchers working within the Chemistry Education Research Group (CERG) at Monash University were provided a random selection of 10% of all responses to Q1 and Q2a/Q2b. They were asked to identify key words suggesting emergent themes in each question and from these emergent themes 'codes' were generated by the primary researcher for participants' responses (Bryman and Burgess, 1994). Having reviewed the data once, the responses were studied in greater

detail to determine whether there were any hidden themes which the initial analysis failed to identify. A third review of the emergent themes within each question was conducted and using a redundancy approach similar themes were combined. This resulted in 21 unique themes for Q1 and 19 unique themes for Q2a/Q2b to used in coding all responses.

The data from the emergent themes of each question was then analysed quantitatively. To determine the number of participants within each group describing a specific theme, the total number of responses within each theme per group was determined using Nvivo's 'Matrix Coding' function. This data was exported to Microsoft Excel and the number of participants describing a specific theme within each group was then expressed as a percentage. This percentage was determined using the number of responses for a theme within a group divided by the total number of participants who answered a given question from that group. These percentages were then presented graphically.

Results

470 students, 106 teaching staff and 43 employers responded to the question: 'What does the term "Critical Thinking" mean to you?' (Q1). 410 of these students and 86 of the teaching staff also responded to the question 'Can you provide an example of when you have had the opportunity to develop your critical thinking while studying chemistry?' (Q2a) or 'Can you provide an example of when you have provided students with the opportunity to develop their critical thinking while studying chemistry?' (Q2b) in the case of the teaching staff.

Table 1 shows the gender distribution and median age of students who chose to provide this data. As can be seen, there is a slightly larger population of male students, by 12%. The median age of students is 19 years old which is the typical age of most first or second year Australian undergraduate university students.

Table 2 shows the teaching activities and year levels taught by the various cohorts within the teaching staff group. Respondents were able to select multiple teaching activities and year levels taught. The TA cohort typically taught first year laboratory sessions whereas senior TAs and academics all taught at various year levels *via* laboratory, tutorial and lecture activities.

Table 3 provides the demographic data for employers. The respondents' main offices were predominantly found in Australia and the respondents themselves generally had a tertiary level qualification, with 40% of respondents holding

Table 1 Demographic data of all undergraduate student participants

Student cohort	Gender with which the students identify		Median student age
	Male	Female	
1st year students	53% (<i>n</i> = 151)	47% (<i>n</i> = 132)	18 (<i>n</i> = 216)
2nd year students	59% (<i>n</i> = 107)	41% (<i>n</i> = 73)	19 (<i>n</i> = 129)
3rd year students	57% (<i>n</i> = 13)	43% (<i>n</i> = 10)	20 (<i>n</i> = 18)
All undergraduates	56% (<i>n</i> = 271)	44% (<i>n</i> = 215)	19 (<i>n</i> = 363)

Table 2 Teaching activities and year levels taught by respondents

		Teaching staff cohort		
		TAs	Senior TAs/ academics	Online academic
Teaching activity	Laboratory	<i>n</i> = 30	<i>n</i> = 9	<i>n</i> = 46
	Tutorial	<i>n</i> = 2	<i>n</i> = 8	<i>n</i> = 44
	Lectures	<i>n</i> = 0	<i>n</i> = 11	<i>n</i> = 50
Year levels taught	No experience	<i>n</i> = 12	<i>n</i> = 0	<i>n</i> = 0
	1st year	<i>n</i> = 30	<i>n</i> = 6	<i>n</i> = 48
	2nd year	<i>n</i> = 8	<i>n</i> = 10	<i>n</i> = 44
	3rd year	<i>n</i> = 1	<i>n</i> = 10	<i>n</i> = 44
	Hons/M/PhD	<i>n</i> = 0	<i>n</i> = 3	<i>n</i> = 48

Total participants per cohort: TAs (*n* = 40), senior TAs/academics (*n* = 12), online academics (*n* = 54).

a PhD. The most common sector in which respondents worked were chemical, pharmaceutical or petrochemicals (16%). There was also a reasonable representation of respondents from development, innovation or manufacturing (12%), life sciences (14%) and government (12%).

The 21 themes generated in response to the question: 'What does the term "Critical Thinking" mean to you?' (Q1) can be found in Table 4 along with a definition and brief quote to illustrate the meaning attributed to these themes. The quantitative analysis found in Fig. 1 describes the frequency with which each of these themes was expressed by students, teaching staff and employers.

It is important to note that a single response may be coded to multiple themes or in some instances none at all. Table 5 provides a breakdown of how many responses contain a given number of themes. For example 87 responses from the first year cohort contain only a single theme whereas 11 responses from employers contain three themes. The mean number of themes per response or coding density was determined for each cohort and each group. Students described a mean value of 1.73 themes per response, teaching staff described an average of 2.75 themes per response and employers described 3.98 themes per response.

In response to the question; 'Can you provide an example of when you have had the opportunity to develop your critical thinking while studying chemistry?' (Q2a) or 'Can you provide an example of when you have provided students with the opportunity to develop their critical thinking while studying chemistry?' (Q2b) 19 themes were generated. Table 6 contains these themes, their definitions and brief excerpts to convey the meaning attributed to these themes. The quantitative analysis found in Fig. 2 describes the frequency with which each of these themes was expressed in student and teaching staff responses.

Once again a single response could be coded to multiple themes or none at all. Table 7 shows how many responses contained a given number of themes. For example 108 first year responses were coded to a single theme compared to only two Senior TA/Academic responses. Students described an average of 1.32 themes per response and teaching staff described an average of 2.25 themes per response.

Discussion

This study aimed to collect chemistry undergraduate students', teaching staff and employers' views pertaining to their definition of critical thinking and in the case of the students and teaching staff, where they believed critical thinking was developed when studying chemistry at university. Many clear patterns emerged from the qualitative analysis. However, the representation and limitations of the data set must be considered before making any generalisations.

Data representation and limitations

All student questionnaires were conducted in a laboratory environment whereas teaching staff completed the questionnaire in a number of environments; in the laboratory, in their office or online. This may have impacted the number of themes per response or coding density of each cohort for a given question. Table 5 illustrates that when responding to Q1 students typical described at least one theme whereas teaching staff and

Table 3 Demographic data of employer participants: country of main office, industry sector and highest qualification held

Country		Sector		Qualification	
Australia	72% (<i>n</i> = 31)	Chemical ^b	16% (<i>n</i> = 7)	PhD	40% (<i>n</i> = 17)
UK	26% (<i>n</i> = 11)	Development ^c	12% (<i>n</i> = 5)	Masters	21% (<i>n</i> = 9)
Belgium	2% (<i>n</i> = 1)	Science ^d	14% (<i>n</i> = 6)	Grad. dip.	5% (<i>n</i> = 2)
		Government	12% (<i>n</i> = 5)	Post-grad cert.	2% (<i>n</i> = 1)
		Health ^e	9% (<i>n</i> = 4)	Bachelors	30% (<i>n</i> = 13)
		Environment ^f	7% (<i>n</i> = 2)	High school	2% (<i>n</i> = 1)
		FMCG ^g	7% (<i>n</i> = 2)		
		Mining	5% (<i>n</i> = 2)		
		Consulting	5% (<i>n</i> = 2)		
		Education	5% (<i>n</i> = 2)		
		Chemical and Development ^a	5% (<i>n</i> = 2)		
		Chemical and FMCG ^a	2% (<i>n</i> = 1)		
		Government and Environment ^a	2% (<i>n</i> = 1)		
		Other	5% (<i>n</i> = 2)		

^a These employers identified multiple sectors and were thus coded according to both themes. ^b Chemical, pharmaceuticals or petrochemicals.

^c Development, innovation or manufacturing. ^d Science or life-science. ^e Health, medical or pathology. ^f Environment or conservation. ^g Fast moving consumer goods.

Table 4 Themes emerging in responses to Q1

Theme	Definition	Example
Analysis	Information, data or evidence analysed or broken down.	"Ability to unpack complex situations..."
Application of knowledge	What is known or learnt is applied in some way.	"Evaluate...from first principles and personal knowledge..."
Arriving at an outcome	The end product of critical thinking. <i>E.g.</i> conclusion, argument or course of action.	"...form a valid, informed opinion." "...an appropriate solution."
Context (macro)	Implication of an outcome with much greater boundaries at an organisational or societal level.	"Ethical and economical solution." "Outside aspects and factors..."
Creative	'Creative thinking' or discussed innovation.	"...imaginative generation of ideas."
Critique	Identify assumptions, reasoning, arguments or presumed facts and determining credibility, validity and reliability.	"...question the concepts..." "...challenging the evidence..."
Decision making	Used in 'making a decision' or for example 'arriving at a decision'.	"...make an informed decision..."
Evaluate	Attributing value to a stimulus. Appraising, determining value or identifying meaning.	"Reflecting on the meaning..." "...filtering of that info..."
Identification of opportunities and problems	Appropriate questioning to understand a problem. Identification of potential issues or opportunities.	"identify where intervention will have the most impact"
Interpretation of information	Engaging with a stimulus and understanding that information.	"...interpreting the data..." "...understand concepts..."
Lateral thinking	Use of the term 'lateral thought' and 'out of the box'.	"... (Thinking) in an abstract manner."
Logical approach	Application of a logic, reasonable or rational thought process.	"...reasoned judgements..." "...finding a rational truth..."
Objectivity	Taking an unbiased approach. Sceptical or open minded.	"...consider various points of view..."
Problem solving	Problem and/or something that needs to be resolved.	"...work through a problem..."
Productivity	Thinking which in some way has a constructive use, <i>e.g.</i> efficient.	"...where intervention will have the most impact..."
Reflection	Metacognitive processes of 'why am I thinking what I'm thinking?'	"Thinking about your thinking"
Research	Collection of (experimental) data, evidence or information.	"...gathering information or data..."
Systematic approach	How thoughts are organised. Order of operations.	"...arrange it (information) in a way that it informs outcomes."
Testing	Exploring and testing knowledge, evidence, claims or arguments.	"...draw conclusions based on hypothesis testing."
Under pressure	Time constraint or when stakes are high.	"...under pressure situations..."
Understanding the local context	Action or opinion is required and have some sort of impact.	"...what must be done in a situation..."

employers on average described at least 3 themes. This lower coding density may be due to maturity or possibly less exposure to critical thinking activities to be able to articulate what it is. Alternatively, since students typically received this questionnaire at the end of a two to three hour laboratory activity they may have been less inclined to write lengthy responses.

A similar pattern of coding density can be observed between TAs (D) *versus* Senior TAs and Academics (E), Online Academics (F) and Employers (G). It would appear that those participants who were approached directly or online made a concerted effort to respond to the questions as can be seen in Tables 5 and 7,

where at least 3 themes were typically described by cohorts E, F and G. Again it is worth considering the experience that cohort D have with critical thinking. The majority of this cohort were on semester long contracts and only had teaching experience in a first year laboratory environment (Table 2). It is possible these participants may not exercise their critical thinking skills as frequently as academics who routinely engage in activities such as peer reviewing journal submissions which exercise these skills more frequently. This aligns with the constructivist notion that an individual creates their meaning of a given construct from their environment (Lemanski and Overton, 2011)

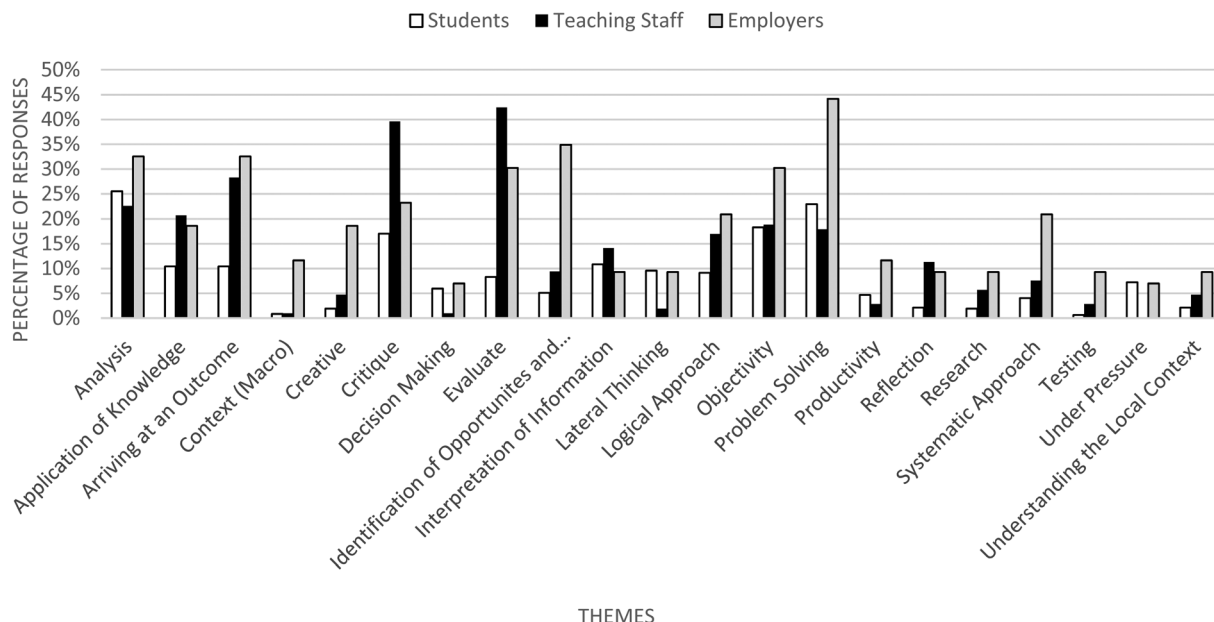


Fig. 1 Percentage of responses describing a given theme by cohort in response to Q1.

Table 5 Number of responses describing a given number of themes in response to Q1

No. themes	0 ^a	1	2	3	4	5	6	7	8	9	
Cohort/group	Number of responses which described the above number of themes										Mean
1st year	30	87	90	37	19	2	1	1			1.79
2nd year	32	55	51	28	12	1					1.64
3rd year	5	6	7	3	3						1.71
Students	67	148	148	68	34	3	1	1			1.73
TAs	5	10	15	7	3						1.83
Sen TAs/academics	1		2	4	5						3.00
Online		4	12	13	17	6	2				3.28
Teaching staff	6	11	26	21	24	6	2				2.75
Employers		1	8	11	8	7	4	3		1	3.98

^a Responses that were coded to zero themes were either considered not to make sense or responses that could not be given meaning without further investigation (see discussion).

and in this research how the participants believe that construct is applied in their daily lives.

With respect to demographic data, there was a slightly larger representation of students identifying as male compared to female. This was observed in all student cohorts, however it is important to note that there was slightly larger number of female students enrolled in chemistry at Monash University as compared to male students. As can be seen from Table 1, the median age for students was nineteen years old. This value was skewed slightly as a result of such large numbers of respondents from first and second year cohorts.

Larger samples of first and second year students and first year TAs were obtained due to the environments in which the questionnaire was conducted (namely compulsory laboratory sessions). Aside from the slightly larger number of male student respondents, there can be some confidence that the data obtained is representative of a random

sample of the respective cohorts and the findings may be generalizable.

Obtaining data from senior TAs, academics and employers was far more difficult and consequently the data collected was more reflective of non-representative convenience sampling. Therefore, the findings herein may have limited generalisability with respect to senior TAs, academics and employers.

Defining critical thinking

As can be seen from Table 5, over sixty responses were attributed no themes. This was for one of two reasons. Respondents may have attempted to demonstrate their wit rather than their understanding of the term critical thinking with responses such as “means a lot”, “not annoying your TA” or “thinking critically”. More commonly responses that were attributed no themes were due to defining critical thinking as “thinking on a much deeper level” or “thinking in a complex manner”.

Table 6 Themes emerging from responses to Q2a/Q2b

Theme	Definition	Example
Algorithmic problem solving	All the data is provided and the solutions are known.	"Solving chemical equations."
Application of knowledge	Use of knowledge, usually developed within a course.	"...fundamental knowledge..."
Assessing knowledge	Formative or summative feedback assessments.	"Weekly tests & semester exams."
Creating an argument	Generate hypothesis, opinion, argument or conclusion.	"...justify their (students) choices..."
Critiquing	Decide quality of experimental data, method or argument.	"...assess aspects of experimental design."
Developing knowledge	Developing specific content knowledge.	"...perform lab task & understand the theory..."
Discussion method	Engaged in dialogue with students, TAs or academics.	"...what would happen if...?"
Engaging with experimental data	Engaging with data generated in the lab or from research.	"...did not achieve expected results..."
Experimental design	Developing an experimental method in a laboratory setting.	"...create their (students) own experiments."
Inquiry based learning	Often described as an 'IDEA prac'.	"...we had to come up with a method for a prac"
Leading questions	Thought processes guided by open-ended questions.	"...prompt them (students) with questions..."
Lecture environment	Activity taking place in a lecture or as part of a lecture.	"Reading materials before and after lectures..."
Open-ended problem solving	Not all the data provided, ill defined or unknown solution.	"...direct answers may not be able to be found..."
Practical environment	Activities taking place in/or as result of a laboratory.	"...making paracetamol..."
Project work	Project which occurred over an extended time period.	"...problem to solve for the semester..."
Research	Use of the term 'research' or indicating research	"...research projects..."
Safety	Safe procedures in a laboratory environment.	"...handling dangerous chemicals."
Testing	Experimentation or test of a hypothesis.	"Not just assuming our hypothesis is right."
Tutorial environment	Tutorials, 'tutes' or 'tutorial questions'.	"...through discussions in tutorial."
Writing	Lab reports, essays or literature reviews.	"Writing lab reports with discussion (section)."

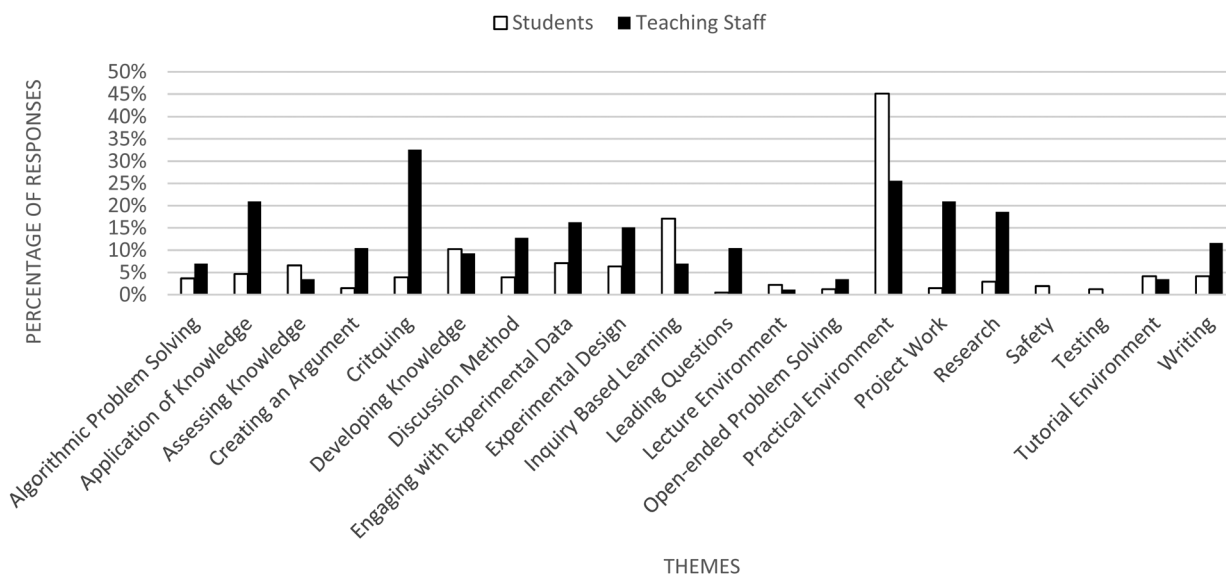


Fig. 2 Percentage of responses describing a given theme by cohort in response to Q2a/Q2b.

Table 7 Number of responses describing a given number of themes for Q2a/Q2b

No. themes	0 ^a	1	2	3	4	5	6	
Cohort/group	Number of responses which described the above number of themes							Mean
1st year	53	108	42	10	1			1.06
2nd year	7	70	89	7	1			1.57
3rd year	0	6	4	2	3			2.13
Students	60	184	135	19	5			1.32
TAs	3	9	11	3				1.54
Sen TAs/acad.	0	2	5	2	2			2.36
Online acad.	0	14	11	16	6	4	1	2.58
Teaching staff	3	25	27	21	8	4	1	2.25

^a Responses that were coded to zero themes were done so as these responses were either not to make sense, responses that could not be given meaning without further investigation, or stated that the respondent had either not studied or taught chemistry before (see Discussion).

The meaning of these responses was unclear and difficult to code. Whilst it could be argued that many themes, such as 'analysis' would benefit from probing *via* focus groups (Edmunds and Brown, 2013, pp. 22–24), this is particularly true of terms such as 'deep thinking'. Consequently, these responses were not coded.

The theme 'analysis' was frequently expressed by all groups (students, teaching staff and employers). At least 20% of all responses identified analysis as part of the meaning of critical thinking. In the case of the student group it was, in fact, the most common theme, with just over 25% of respondents using it to define critical thinking. The term analysis or analysing was commonly used to describe interaction with some sort intellectual stimulus, whether it be an idea, data or a problem. Many responses referred to 'analysing something' to suggest a breath of critical thinking.

Students strongly identified with three other themes: 'critique', 'objectivity' and 'problem solving'. Problem solving was the second most commonly expressed theme by student respondents with just over 23% of responses describing it. The link between critical thinking and problem solving appears to be a common association made by students (Tapper, 2004). Critique and objectivity were identified in approximately 17% of responses. The relatively smaller number of themes described by students is not altogether surprising as other qualitative studies have shown students often have difficulty conceptualising critical thinking (Duro *et al.*, 2013).

Teaching staff most commonly described the themes 'critique' (40%) and 'evaluate' (42%) when defining critical thinking. In other recent studies a similar emphasis on interpreting information *via* analysis and evaluation was also observed (Duro *et al.*, 2013; Desai *et al.*, 2016). Teaching staff were much more goal orientated than students with 28% of responses describing 'arriving at an outcome'. Outcomes were very task orientated a kin to Barnett's (1997) 'critical being', either developing a plan relating to experimental design or arriving at a conclusion as a result of experimental data. For example:

"The ability to examine evidence, come to a conclusion based on that evidence..."

Teaching staff also commonly described the themes 'application of knowledge', 'logical approach', 'objectivity' and 'problem solving' in approximately 20% of responses. It is worth noting that

students and teaching staff express the theme of 'objectivity' with similar frequencies (18% and 19%, respectively). Of all three groups, teaching staff use the theme of problem solving the least when defining critical thinking (18%). While only 14% of teaching staff respondents described the theme of 'interpreting information' the value of this as being part of critical thinking was higher than with the student (11%) and employer (9%) groups.

As can be seen from Table 5 employers typically described the largest number themes in their responses. 'Problem solving' was the most common theme expressed by over 44% of employers. Employers were goal orientated much like teaching staff, commonly describing themes of 'application of knowledge' (19%), 'objectivity' (30%), 'logical approach' (21%), 'evaluate' (30%) and 'arriving at an outcome' (33%). Arriving at an outcome contained a wide breadth of examples in employer responses. However, there was some focus on using evidence to inform a conclusion which would lead to a course of action for the organisation to take:

"...a necessary approach to solving or answering problems, developing a product or process."

Employers expressed four themes unique to their group: 'context (macro)' (12%), 'creative' (19%), 'systematic approach' (21%) and 'identification of opportunities and problems' (35%). The latter focused on the use of critical thinking as a method of uncovering what is not immediately apparent:

"To consider the problem to expose route cause(s) in a rationale and logical manner and apply lateral thinking to seek solutions to the problem."

The above response also includes in its definition of critical thinking;

"The ability of a person to identify a problem that does not have a readily available or off the shelf solution."

This is an excellent example of responses identifying creativity in conjunction with the theme of problem identification. The general sentiment of employers was that critical thinking is important to innovation within the organisation and is suggestive of what Jackson (2010b) refers to as 'Pro-c creativity' or the creativity associated within a professional environment.

Furthermore, employers were unique in describing critical thinking with the theme of 'context (macro)'. What this theme references is that employers identified the application of critical thinking on a much broader social scale. For example:

"...understand the implications from an organisational perspective."

"...collaborating the thoughts and views of others to gain a clearer insight of the real challenge."

Employers acknowledged that the results of critical thinking can have an impact in commercial and societal contexts. While students and teaching staff have a somewhat more internalised definition of critical thinking, employers appear to have a more social application of critical thinking as seen in some the literature (Desai *et al.*, 2016).

One of the most interesting features of this data was that the terms 'judgement' and 'inference', found in the Delphi definition of critical thinking (Facione, 1990), were seldom used by respondents. In fact below are the only two student responses to use the term 'judgement':

"Not taking things at face value and giving topics considerable thought and analysis before coming to a conclusion/judgement on it." – First year respondent

"Analysis of a problem to make a judgement." – Second year respondent

It is worth noting that a similar minority of respondents used the term 'opinion' in their definition of critical thinking;

"Ability to objectively analyse, process and form an opinion of a particular subject."

And a slightly larger number of respondents used the term 'conclusion':

"A skill to understand a thing more clearly and make conclusion."

When the Delphi report describes core critical thinking skills the terms 'judgement' and 'opinion' are used somewhat synonymously. Similarly, 'drawing conclusions' is explicitly stated as a sub skill of the skill of 'inference' (Facione, 1990, p. 10). This suggests that a larger number of respondents using 'opinion' or 'conclusion' may in fact be referencing the terms 'judgement' or 'inference'. However without further probing what respondents mean by 'conclusion' or 'opinion' this is not a certainty.

There is also very little emphasis around self-regulation or the metacognitive processes typically associated with 'good' critical thinking (Glaser, 1984; Bailin, 2002). Perhaps this is implied when respondents described the theme of 'objectivity':

"Thinking about situations with an open view point and analysing what you're doing."

What is very clear from this data is the emphasis on problem solving in the definition of critical thinking. This was a very prominent feature of the data from students and employers. With respect to the students this may be due to the perception that

scientific facts are unquestionable and the algorithmic problem solving pedagogies commonly employed in science education (Zielinski, 2004; DeWit, 2006; Cloonan and Hutchinson, 2011). This feature of the data was slightly less common in teaching staff, but it was very prominent with employers. This might be due to the fact that employers are typically adept at reflecting on open-ended problems and identifying any parameters or approximations required (Randles and Overton, 2015). This experience with open-ended problems may also explain the description of the theme of 'identification of problems and opportunities' which was somewhat unique to employers.

Interestingly the Delphi report does not consider problem solving an element of critical thinking. Instead it proposes problem solving and critical thinking are 'closely related forms of higher-order thinking' (Facione, 1990, p. 5). Similarly Halpern suggests that certain behaviours are associated with critical thinking or problem solving but that these higher order cognitive skills are not mutually exclusive (Halpern, 1996, pp. 317–363). This cognitive psychology view is more reflective of the data that has emerged from respondents in this study which might otherwise be considered misconceptions with respect to critical thinking.

Regardless of this interpretation, it would be interesting to ask students, teachers and professionals from other disciplines to define critical thinking. It is quite possible that an emphasis on judgement may occur in humanities, commerce or arts and perhaps there would be less use of the theme of problem solving. For example when a group of business academics were asked to describe which critical thinking skills were important to graduates entering the workforce within their discipline, 47% of responses described problem solving and 34% of responses described analysis (Desai *et al.*, 2016).

The other interesting feature of this data are the points of difference between groups and what these may be attributed to. For example teaching staff emphasised the themes of 'critique' and 'evaluate'. A common aspect of an academics role is to be involved in peer review and academic writing so it is not surprising that these themes arise so frequently. Likewise employers' frequency of themes around identification, innovation and context are reflective of a competitive commercial environment. Given the respondents association between critical thinking and problem solving, these perceptions around evaluation and identifying problems could also be a reflection of behaviours typical of expert open-ended problem solvers (Randles and Overton, 2015). Both employers and teaching staff have a goal oriented definition of critical thinking which may be a product of maturity and/or their exposure to professional environments. Again this may be an example of constructivism (Lemanski and Overton, 2011).

As can be seen in Table 8, all groups used themes around analysis, critiquing, objectivity and problem solving to define critical thinking. In addition teaching staff and employers use themes relating to the application of knowledge, arriving at an outcome, evaluation and using a logical approach. Employers further expand on their definition to include themes regarding creativity, considering the broader context, taking a systematic approach and identifying opportunities and problems. These

Table 8 Common themes emerging by group in responses to Q1

Theme	Students (%)	Teaching staff (%)	Employers (%)
Analysis	> 20	> 20	> 30
Application of knowledge		> 20	> 10
Arriving at an outcome		> 20	> 30
Context (macro)			> 10
Creative			> 10
Critique	> 10	> 40	> 20
Evaluation		> 40	> 20
Identifying opportunities...			> 30
Logical approach		> 10	> 20
Objectivity	> 10	> 10	> 20
Problem solving	> 20	> 10	> 40
Systematic approach			> 20

Expressed as increments of >10% for ease of readability and to highlight similarities and differences between groups.

themes regarding the definition of critical thinking can be synthesised thus:

To analyse and critique objectively when solving a problem. – Students

To analyse, critique and evaluate through the logical and objective application of knowledge to arrive at an outcome when solving a problem. – Teaching staff

To analyse, critique and evaluate problems and opportunities through the logical, systematic, objective and creative application of knowledge so as to arrive at an outcome and recognise the large scale context in which these problems and opportunities occur. – Employers

While there are some similarities between the definitions of critical thinking it would be inaccurate to suggest that there is a shared definition. Furthermore, the depth to which critical thinking was defined appears to reflect the constructivist phenomena. Employers most commonly reflect definitions found in the literature (Facione, 1990; Halpern, 1996; Tiruneh *et al.*, 2014). Employers appear to have a broader definition of critical thinking and this may be related to the fact that employers work in very broad contexts and a range of experiences, going beyond chemistry to deal with issues such as budgets, policies and human resources.

Where is critical thinking developed while studying chemistry at university?

Much like Q1 some responses were not attributed to themes (see Table 7). This was predominantly in the student group. While some respondents continued to demonstrate their aptitude for comedy and not seriously engage with the questionnaire, the majority of responses which weren't coded stated they had not previously studied chemistry. Similarly, ten TAs were just commencing their honours research year and had not had any teaching experience at the time of the questionnaire.

With respect to the teaching staff, the wording of the question they received must be considered to put the responses in context: 'Can you provide an example of when you have provided students with the opportunity to develop their critical

thinking while studying chemistry?' (Q2b) This wording elicited responses which were drawn from the respondents' recent teaching activities and may actually differ from where the respondent believes students develop their critical thinking most. For example many TAs from cohort A only have practical experience to draw on whereas cohorts B and C also have lecture and/or tutorial activities to base their response on (Table 2). Conversely some respondents from cohorts B and C only had lecture or tutorial experience to draw on.

When asked to provide an example of where they believed they developed their critical thinking while studying chemistry, 45% of students identified an activity relating to a practical environment. The second most common theme was 'inquiry based learning' (17%). What was most interesting was that 36% of second year students and 14% of third year students specifically mentioned 'IDEA pracs'. These practicals were guided inquiry activities the students performed as part of their first year laboratory program (Rayner *et al.*, 2013). The fact that after two years in some cases students identified these activities demonstrates the effectiveness of inquiry-based learning in developing transferable skills such as critical thinking.

It is important to recognise that students do not identify activities that make the teaching of critical thinking explicit. Students in other studies identified courses around scientific communication as opportunities where critical thinking was explicitly taught (Tapper, 2004). Beyond these courses, much like the students in the current study, the development of critical thinking became more implicit and students became dependent on feedback from writing activities (Tapper, 2004; Duro *et al.*, 2013). It is clear from the literature, without a deliberate effort to make critical thinking goals explicit in discipline specific courses, students find it difficult to conceptualise, and perceive critical thinking as an intuitive skill that develops over time (Tapper, 2004; Beachboard and Beachboard, 2010; Duro *et al.*, 2013; Loes *et al.*, 2015).

Teaching staff also identified practical environments (26%) as to when they developed students' critical thinking. However, four additional themes were also prominent in their responses: 'application of knowledge' (21%), 'critique' (33%), 'project work' (21%) and 'research' (19%). These themes are reflective of activities described in recent literature designed to elicit higher order cognitive skills (Cowden and Santiago, 2016; Stephenson and Sadler-Mcknight, 2016; Toledo and Dubas, 2016). Critique activities ranged from critiquing experimental design to writing literature reviews:

"I may provide students with some experimental evidence and they need to evaluate whether these are consistent with specific mechanisms."

"Choosing and researching a topic to conduct a literature review on. Writing a review to include critical appraisal of the information covered."

"Research paper-based assessments in which students are asked to locate and extract information, analyse data and critically assess aspects of experimental design."

“...paper analysis which requires use of many variables in understanding change factors and outcomes in reaction.”

The ‘application of knowledge’ most often described activities taking place predominantly in a lecture environment and in some instances in a practical environment. Themes of ‘project work’ and ‘research’ often described activities in practical environments. Many of these responses focus on final year research projects:

“Mainly this comes from the crucial role of the research project, generally in the final year of study when the student has had the opportunity to build up their knowledge base across a broad range of chemistry.”

The above statement would suggest that critical thinking can only be achieved with a solid foundation of discipline specific knowledge. While it holds true that an individual is a better critical thinker within their discipline specific knowledge (McPeak, 1981; Moore, 2011) it is not true that a large body knowledge is a necessary prerequisite to develop critical thinking (Ennis, 1989; Davies, 2013).

According to this data students and teaching staff have some limited agreement that critical thinking is developed in a practical environment. However, that is where the similarities end. Despite teaching staff believing that they develop critical thinking through the application of knowledge this is not apparent to the students.

Implications for practice

The methods described by teaching staff is what Ennis (1989) described as the Immersion approach, whereby subject matter is covered in great depth but the critical thinking goals are implicit. It would appear the more overt approaches suggested by Ennis (1989) and McMillan (1987) would assist students in recognising when they are being taught critical thinking. This could also contribute to students more thoroughly articulating what critical thinking is.

Teaching staff commonly acknowledge that students develop their critical thinking in active environments in accordance with the literature (Biggs, 2012). However the research projects the respondents commonly describe are often elective subjects or offered as vacation internships, the numbers of which are limited and will only become scarcer as student numbers continue to grow. It would be useful to determine if teaching staff believed project work is an opportunity to measure student critical thinking or whether it is better measured *via* other activities (if at all) and compare this to the literature (Desai *et al.*, 2016).

A recent meta-analysis would suggest, a combination of teaching activities afford the greatest effect with respect to the development of critical thinking (Abrami *et al.*, 2015). These teaching activities according to Abrami and colleagues are described as ‘authentic instruction’, ‘dialogue’ and ‘mentoring’. These findings are reflective of the present work where practical inquiry based learning, discussions and research projects were commonly described as opportunities to develop critical thinking.

It is advisable for chemistry educators wishing to develop critical thinking in students that the activities described by students and teaching staff within this research form a foundation within their practice, emphasising authentic problem solving and Socratic dialogue (Abrami *et al.*, 2015).

Future work

As described earlier, there are several limitations to this study. To further understand the meaning behind terms such ‘deep thinking’ or ‘out-of-the-box’, focus group interviews would prove useful. A larger sample size, particular with respect to third year students within Monash University, teaching staff and employers could improve the quality of the data. The expression of certain themes may become more or less prominent in a larger sample size and would refine the definition of critical thinking described by employers in particular. Likewise, providing this questionnaire to students in other faculties or even other institutions from various countries would add robustness to the findings as the majority of participants were Australian. Those interested conducting the questionnaire are encouraged to contact the authors *via* email.

Conclusion

When looking at the results of this study there are several clear differences between students, teaching staff and employers. These differences may arise from several factors such as education, maturity, experience, environment or possible a combination of all of these. This may be a reflection of the constructivist notion that an individual creates meaning of constructs such as critical thinking as result of and through interacting with their environment. This is exemplified by the emphasis on problem solving by students and employers when defining critical thinking whereas teaching staff more commonly associate critiquing and evaluation with critical thinking. Specifically, employers appeared to have a more thorough definition of critical thinking which may be due to broader contexts found in the workplace.

When asked to define critical thinking *via* an open ended questionnaire students, teaching staff and employers all described the themes of analysis, critique, objectivity and problem solving. Teaching staff and employers commonly expressed themes around evaluation, goal orientation and use of logic. Employers also believed creativity, larger scale contexts, taking a systematic approach and identifying of opportunities and problems are important aspects of critical thinking. This would suggest there is only a limited shared definition of critical thinking between students, teaching staff and employers which centres on analysis and problem solving.

In the same open ended questionnaire students and teaching staff described where they believed they developed student critical thinking. Overwhelmingly students described practical environments and inquiry based learning activities developed critical thinking. Teaching staff expressed themes around the application and critiquing of knowledge and to some extent practical environments and research projects. Again there

appeared to be limited overlap between the perceptions of students and teaching staff and the need for more immersive student experiences, such as inquiry-based learning and work integrated learning (Edwards *et al.*, 2015), is apparent in the development of transferable skills such as critical thinking.

If the workplace is expecting tertiary institutes to provide chemistry graduates for the workforce, a shared definition of critical thinking is imperative. However, there appears to be a somewhat limited shared understanding as to what critical thinking skills entail. If there are so many facets to critical thinking how can universities accommodate the development of these? Initiatives such work integrated learning (Edwards *et al.*, 2015) aim to give students experience in commercial environments and perhaps in combination with inquiry-based pedagogies, a shared understanding of critical thinking and how to develop it can occur.

Acknowledgements

The authors would like to acknowledge undergraduate and teaching staff participants from Monash University and academics and employers who took the time to complete the questionnaire online. This research was made possible through the Australian Post-graduate Award funding and with guidance of the Monash University Human Ethics Research Committee.

References

- Abrami P. C., Bernard R. M., Borokhovski E., Waddington D. I., Wade C. A. and Persson T., (2015), Strategies for teaching students to think critically: a meta-analysis, *Rev. Educ. Res.*, **85**(2), 275–314.
- Australian National University, (2015), *Chemistry major*, retrieved from <http://programsandcourses.anu.edu.au/2016/major/CHEM-MAJ>.
- Bailin S., (2002), Critical thinking and science education, *Sci. Educ.*, **11**, 361–375.
- Barnett R., (1997), *Higher education: a critical business*, Buckingham: Open University Press.
- Beachboard M. R. and Beachboard J. C., (2010), Critical-thinking pedagogy and student perceptions of university contributions to their academic development. (report), *Informing Science: the International Journal of an Emerging Transdiscipline*, **13**, 53–71.
- Biggs J., (2012), What the student does: teaching for enhanced learning, *Higher Educ. Res. Dev.*, **31**(1), 39–55.
- Bryman A. and Burgess R. G., (1994), Reflections on qualitative data analysis, in Bryman A. and Burgess R. G. (ed.), *Analyzing qualitative data*, London: Sage.
- Cloonan C. A. and Hutchinson J. S., (2011), A chemistry concept reasoning test, *Chem. Educ. Res. Pract.*, **12**, 205–209.
- Cowden C. D. and Santiago M. F., (2016), Interdisciplinary explorations: promoting critical thinking via problem-based learning in an advanced biochemistry class, *J. Chem. Educ.*, **93**, 464–469.
- Davies M., (2013), Critical thinking and the disciplines reconsidered, *Higher Educ. Res. Dev.*, **32**(4), 529–544.
- Desai M. S., Berger B. D. and Higgs R., (2016), Critical thinking skills for business school graduates as demanded by employers: a strategic perspective and recommendations, *Academy of Educational Leadership Journal*, **20**(1), 10–31.
- DeWit D. G., (2006), Predicting inorganic reaction products: a critical thinking exercise in general chemistry, *J. Chem. Educ.*, **83**, 1625–1628.
- Dressel P. L. and Mayhew L. B., (1954), *General education: explorations in evaluation*, Washington, D.C.: American Council on Education.
- Duro E., Elander J., Maratos F. A., Stuppel E. J. N. and Aubeeluck A., (2013), In search of critical thinking in psychology: an exploration of student and lecturer understandings in higher education, *Psychology Learning & Teaching*, **12**, 275–281.
- Edmunds S. and Brown G., (2013), Section 6: undertaking pedagogic research using qualitative methods, in Groves M. and Overton T. (ed.), *Getting started in pedagogic research within the stem disciplines*, Edgbaston, Birmingham, UK: University of Birmingham STEM Education Centre, pp. 21–26.
- Edwards D., Perkins K., Pearce J. and Hong J., (2015), Work integrated learning in stem in Australian universities: Australian Council for Education Research.
- Ennis R. H., (1989), Critical thinking and subject specificity: clarification and needed research, *Educ. Res.*, **18**(3), 4–10.
- Ennis R. H., (1990), The extent to which critical thinking is subject-specific: further clarification, *Educ. Res.*, **19**(4), 13–16.
- Facione P. A., (1990), Critical thinking: a statement of expert consensus for purposes of educational assessment and instruction, *Executive summary. "The delphi report"*. Millbrae, CA.
- Flynn A. B., (2011), Developing problem-solving skills through retrosynthetic analysis and clickers in organic chemistry, *J. Chem. Educ.*, **88**, 1496–1500.
- Ghulam R. N. and David B., (1999), Graduates' perceptions of transferable personal skills and future career preparation in the UK, *Education + Training*, **41**(4), 184–193.
- Glaser R., (1984), Education and thinking: the role of knowledge, *Am. Psychol.*, **39**(2), 93–104.
- Graduate Careers Australia, (2015), *Chemistry – bachelor graduates (all)*, retrieved from <http://www.graduatecareers.com.au/Research/GradJobsDollars/BachelorAll/Chemistry/index.htm>.
- Gupta T., Burke K. A., Mehta A. and Greenbowe T. J., (2015), Impact of guided-inquiry-based instruction with a writing and reflection emphasis on chemistry students' critical thinking abilities, *J. Chem. Educ.*, **92**(1), 32–38.
- Halpern D. F., (1993), Assessing the effectiveness of critical thinking instruction, *J. Gen. Educ.*, **50**(4), 238–254.
- Halpern D. F., (1996), *Thought and knowledge: An introduction to critical thinking*, Mahwah, N.J.: L. Erlbaum Associates.
- Halpern D. F., (1998), Teaching critical thinking for transfer across domains. Dispositions, skills, structure training, and metacognitive monitoring, *Am. Psychol.*, **53**, 449–455.

- Henderson D. E., (2010), A chemical instrumentation game for teaching critical thinking and information literacy in instrumental analysis courses, *J. Chem. Educ.*, **87**, 412–415.
- Inhelder B. and Piaget J., (1958), *The growth of logical thinking from childhood to adolescence: an essay on the construction of formal operational structures*, London: Routledge & Kegan Paul.
- Jackson D., (2010a), An international profile of industry-relevant competencies and skill gaps in modern graduates, *International Journal of Management Education*, **8**(3), 29–58.
- Jackson N., (2010b), Developing creativity for professional capability through lifewide education, in Jackson N. (ed.), *Learning to be Professional through a Higher Education*, retrieved from <http://learningtobeprofessional.pbworks.com/f/JACKSON+A4.pdf>.
- Jones S., Yates B. and Kelder J.-A., (2011), *Learning and teaching academic standards project, science: learning and teaching academic standards statement September 2011*, retrieved from http://www.acds-tlcc.edu.au/wp-content/uploads/sites/14/2015/02/altc_standards_SCIENCE_240811_v3_final.pdf.
- Klein G. C. and Carney J. M., (2014), Comprehensive approach to the development of communication and critical thinking: bookend courses for third- and fourth-year chemistry majors, *J. Chem. Educ.*, **91**, 1649–1654.
- Leggett M., Kinnear A., Boyce M. and Bennett I., (2004), Student and staff perceptions of the importance of generic skills in science, *Higher Educ. Res. Dev.*, **23**, 295–312.
- Lehman D. R. and Nisbett R. E., (1990), A longitudinal study of the effects of undergraduate training on reasoning, *Dev. Psychol.*, **26**, 952–960.
- Lehman D. R., Lempert R. O. and Nisbett R. E., (1988), The effects of graduate training on reasoning: formal discipline and thinking about everyday-life events, *Am. Psychol.*, **43**, 431–442.
- Lemanski T. and Overton T., (2011), *An introduction to qualitative research*, retrieved from <https://hydra.hull.ac.uk/assets/hull:4506/content>.
- Loes C. N., Salisbury M. H. and Pascarella E. T., (2015), Student perceptions of effective instruction and the development of critical thinking: a replication and extension, *Higher Education: The International Journal of Higher Education Research*, **69**, 823–838.
- Lowden K., Hall S., Elliot D. and Lewin J., (2011), *Employers' perceptions of the employability skills of new graduates: research commissioned by the edge foundation*, retrieved from http://www.educationandemployers.org/wp-content/uploads/2014/06/employability_skills_as_pdf_-_final_online_version.pdf.
- Mann L., (1979), *On the trail of process: a historical perspective on cognitive processes and their training*, New York: Grune & Stratton.
- Martineau E. and Boisvert L., (2011), Using wikipedia to develop students' critical analysis skills in the undergraduate chemistry curriculum, *J. Chem. Educ.*, **88**, 769–771.
- Matthews M. R., (1993), Constructivism and science education: some epistemological problems, *J. Sci. Educ. Technol.*, **2**(1), 359–370.
- McMillan J., (1987), Enhancing college students' critical thinking: a review of studies, *J. Assoc. Inst. Res.*, **26**(1), 3–29.
- McPeak J. E., (1981), *Critical thinking and education*, Oxford: Martin Roberston.
- Monash University, (2015), *Undergraduate - area of study. Chemistry*, retrieved from <http://www.monash.edu.au/pubs/2015handbooks/aos/chemistry/>.
- Moore T., (2013), Critical thinking: seven definitions in search of a concept, *Studies in Higher Education*, **38**(4), 506–522.
- Moore T. J., (2011), Critical thinking and disciplinary thinking: a continuing debate, *Higher Educ. Res. Dev.*, **30**(3), 261–274.
- Nisbett R. E., Fong G. T., Lehman D. R. and Cheng P. W., (1987), Teaching reasoning, *Science*, **238**, 625–631.
- Oliver-Hoyo M. T., (2003), Designing a written assignment to promote the use of critical thinking skills in an introductory chemistry course, *J. Chem. Educ.*, **80**, 899–903.
- Pithers R. T. and Soden R., (2000), Critical thinking in education: a review, *Educ. Res.*, **42**(3), 237–249.
- Prinsley R. and Baranyai K., (2015), *Stem skills in the workforce: what do employers want?* retrieved from http://www.chiefscientist.gov.au/wp-content/uploads/OPS09_02Mar2015_Web.pdf.
- Randles C. A. and Overton T. L., (2015), Expert vs. novice: approaches used by chemists when solving open-ended problems, *Chem. Educ. Res. Pract.*, **16**(4), 811–823.
- Rayner G. M., Charlton-Robb K.-M., Thompson C. D. and Hughes T., (2013), Interdisciplinary collaboration to integrate inquiry-oriented learning in undergraduate science practicals, *Int. J. Innovation Sci. Math. Educ.*, **21**(5), 1–11.
- Sarkar M., Overton T., Thompson C. and Rayner G., (2016), Graduate employability: views of recent science graduates and employers, *Int. J. Innovation Sci. Math. Educ.*, **24**(3), 31–48.
- Stephenson N. S. and Sadler-Mcknight N. P., (2016), Developing critical thinking skills using the science writing heuristic in the chemistry laboratory, *Chem. Educ. Res. Pract.*, **17**(1), 72–79.
- Tapper J., (2004), Student perceptions of how critical thinking is embedded in a degree program, *Higher Educ. Res. Dev.*, **23**(2), 199–222.
- The University of Adelaide, (2015), *University of Adelaide graduate attributes*, retrieved from <http://www.adelaide.edu.au/learning/strategy/gradattributes/>.
- The University of Melbourne, (2015), *Handbook – chemistry*, retrieved from <https://handbook.unimelb.edu.au/view/2015/!R01-AA-MAJ%2B1007>.
- Thorndike E. L. and Woodworth R. S., (1901a), The influence of improvement in one mental function upon the efficiency of other functions. (I), *Psychol. Rev.*, **8**(3), 247–261.
- Thorndike E. L. and Woodworth R. S., (1901b), The influence of improvement in one mental function upon the efficiency of other functions. II. The estimation of magnitudes, *Psychol. Rev.*, **8**(4), 384–395.
- Thorndike E. L. and Woodworth R. S., (1901c), The influence of improvement in one mental function upon the efficiency of other functions: functions involving attention, observation and discrimination, *Psychol. Rev.*, **8**(6), 553–564.

- Tiruneh D. T., Verburgh A. and Elen J., (2014), Effectiveness of critical thinking instruction in higher education: a systematic review of intervention studies, *Higher Educ. Stud.*, **4**(1), 1–17.
- Toledo S. and Dubas J. M., (2016), Encouraging higher-order thinking in general chemistry by scaffolding student learning using Marzano's taxonomy, *J. Chem. Educ.*, **93**(1), 64–69.
- Wason P. C., (1966), New horizons, in Foss B. (ed.), *Psychology*, Harmondsworth, England: Penguin.
- Wilson K. L., Lizzio A. and Ramsden P., (1997), The development, validation and application of the course experience questionnaire, *Stud. Higher Educ.*, **22**(1), 33–53.
- Zielinski T. J., (2004), Critical thinking in chemistry using symbolic mathematics documents, *J. Chem. Educ.*, **81**(10), 1533.

The research presented in the preceding paper served as an exploratory study to gain insight into students', teaching staff and employers' conceptualisation of critical thinking. The study also aimed to help guide the development of teaching interventions and assessment pertaining to critical thinking as recommended in the literature. Ennis (1993) suggested that critical thinking tests require a clear purpose, and must adequately address logistical issues such as administering with large cohorts of students. It was hoped that the findings from the research presented in the *Chemistry Education Research and Practice* paper would provide an explicit definition of critical thinking which aligned with the definitions of the students, teaching staff and employers and could be the focus of the DOT-CCTT. However, as was previously described (Danczak, Thompson and Overton, 2017), students, teaching staff and employers predominately identified deductive logic elements of critical thinking, such as 'analysis' and 'problem solving', and neglected to describe inductive logic elements, such as 'judgement' or 'inference', typical of the literature on critical thinking (Facione, 1990; Halpern, 1996b).

Students, teaching staff and employers all appeared to define critical thinking slightly differently. None of these groups appeared to define critical thinking in the holistic fashion of the philosophers, cognitive psychologists or education researchers. In fact, very much in line with the constructivist paradigm (Ferguson, 2007), participants seem to have drawn on elements of critical thinking relative to the environments in which they had previously been required to use critical thinking. The most obvious example of this was that of the students applying their critical thinking skills to problem solving in an education setting.

Ennis stated that critical thinking tests need to align with what aspects of critical thinking students are being taught (Ennis, 1993). While the students described problem solving as how they define critical thinking, it does not necessarily mean the students' description is entirely representative of what critical thinking skills they have been taught. Therefore to base a chemistry critical thinking test solely on analysis and problem solving skills would lead to the omission of the assessment of other important aspects of critical thinking. With that in mind, the findings described in the preceding paper assisted the researcher to develop a functional

definition of critical thinking which encompassed the experiences of students, academics and employers but that also recognised that these views were narrow and based on their experiences within a chemistry context.

2.9 Operational Definition of Critical Thinking

The definitions of critical thinking cited by researchers (Facione, 1990; Halpern, 1996b; Lehman *et al.*, 1988) covers a wide range of skills and behaviours. These definitions often imply that to think critically necessitates that all of these skills or behaviours be demonstrated. However, it seems almost impossible that all of these attributes could be observed at a given point in time, let alone assessed (Bailin, 2002; Dressel and Mayhew, 1954). Figure 2. 2 attempts to generate a visual representation of how the operational definition of critical thinking sits within this body of research. This figure is by no means a quantitative comparison between critical thinking definitions.

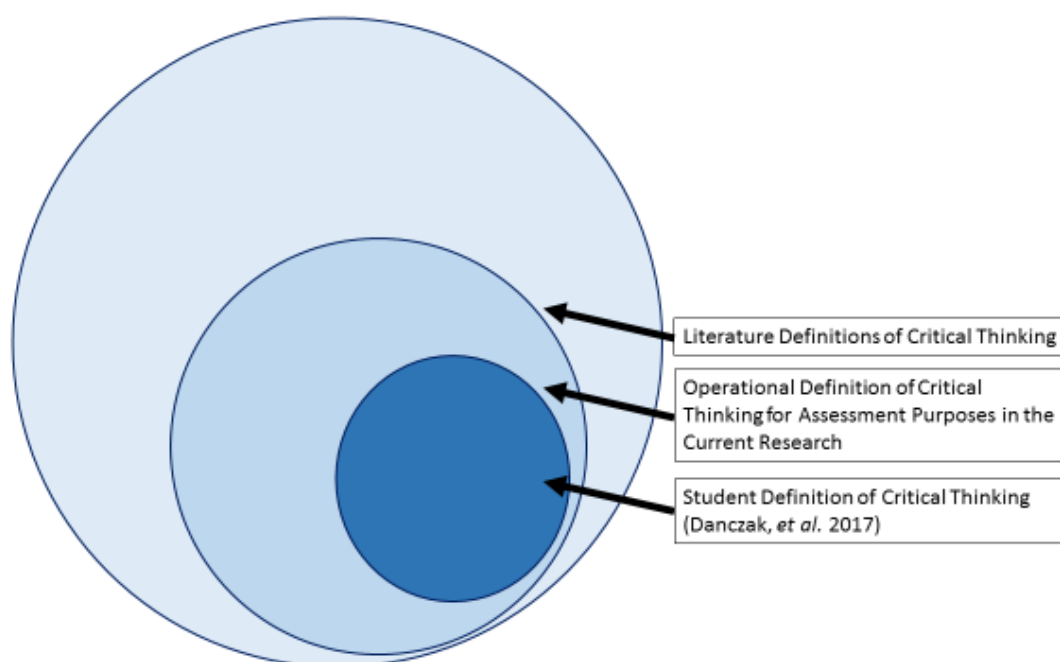


Figure 2. 2 Diagrammatic representation of how the operational definition of critical thinking within this study. The operational definition is situated between the literature definitions of critical thinking and student definition determined by Danczak *et al.* (2017)

Figure 2. 2 serves as a guide to highlight that the definition of critical thinking used throughout this research was broader in context than the definitions presented in Danczak *et al.* (2017), yet not so broad as to encompass every aspect of critical thinking defined in the literature. The largest circle reflects all the definitions of critical thinking, and the skills and dispositions which make it up according to the literature reviewed in Chapter 1. The smallest circle represents the students' narrow perceptions of critical thinking based on the study presented in the journal article above. The circle situated between these two circles highlights the definition of critical thinking used within this thesis. This intermediate position acknowledges the analysis and problem solving focus that students predominately use to describe critical thinking, while expanding into other important aspects of critical thinking such as inference and judgement.

The difference in size between these circles seeks to convey the view that the researcher acknowledges all the elements of critical thinking described in the literature presented as skills or dispositions of critical thinking. However, in light of the research in the journal article presented in Section 2.8, the researcher cannot deny the role that an individual's context plays in their application of critical thinking. As McPeak (1981) suggested; critical thinking is always thinking about something. Environmental and contextual stimuli give rise to certain critical thinking skills and behaviours being required to reach a desirable outcome. At any point in time it is unlikely that an individual can utilise all the skills and behaviours considered hallmarks of a good critical thinker. More realistically, a good critical thinker possesses the meta-cognitive awareness to recognise their environmental and contextual stimuli, and draw on the appropriate critical thinking skills and behaviours to a high degree of competency as the situation dictates. To this end, the researcher recognised that to assess all critical thinking skills and behaviours would be an impossible task, and therefore needed to refine assessment to a sub-set of skill and/or behaviours. The researcher therefore looked to existing critical thinking assessments for guidance, as described below.

2.10 Development of the Danczak-Overton-Thompson Chemistry Critical Thinking Test (DOT-CCTT)

At the time of test development only a privately licenced version of the California Critical Thinking Skills Test (CCTST) (Insight Assessment, 2013) and a practice version of the Watson-Glaser Critical Thinking Appraisal (WGCTA) (AssessmentDay Ltd, 2015) were able to be accessed. As described in Section 1.4, several commercially available critical thinking tests are commonly used in education research and for recruitment purposes. These tests include the CCTST, the Ennis-Weir Critical Thinking Essay Test (EWCTET) (Ennis and Weir, 1985a), the WGCTA, the Cornell Critical Thinking Test Level Z (CCTT-Z) (The Critical Thinking Co., 2017) and the Halpern Critical Thinking Assessment (HCTA) (Halpern, 2016). Several concessions had to be made with respect to the practicality of using commercial products; cost, access to questions and solutions, and reliability of assessment. The WGCTA style questions were chosen as a model for the DOT-CCTT since the practice version of this test was freely available online with accompanying solutions and rationale for those solutions (AssessmentDay Ltd, 2015). The WGCTA is a multiple choice test, and while Ennis (1993) states open-ended response questions more accurately reflect the nature of critical thinking and provide the most reliable results, he conceded that as the number of participants increases the resources, both time and money, impose restrictions and multiple choice questions become more viable.

The WGCTA is one of the oldest critical thinking tests, with its first version dating back to the 1920s and, as such, has undergone the most scrutiny in the literature. It is an 85 item test which covers the core principles of critical thinking: inference, assumption identification, deduction, interpreting information and evaluation of arguments. The test questions within the WGCTA were designed to evaluate the various aspects of critical thinking independent of context. The instructions, parent statements or stems, and the questions themselves were very concise with respect to language and reading requirements. The fact that the WGCTA focused on testing assumptions, deductions, inferences, analysing arguments and interpreting information was an inherent limitation in its ability to assess all critical thinking skills and

behaviours. That said, these elements covered in the WGCTA are skills commonly described by many definitions of critical thinking (Facione, 1990; Halpern, 1996b). Furthermore, given that the assessment of good critical thinking behaviours needs to be readily observable by the assessor, the DOT-CCTT was not intended to make any attempts to measure critical thinking dispositions.

The pilot version of the DOT-CCTT (DOT-CCTTp) was initially developed with 85 questions set within a chemistry or science context using similar instructions and structure to the WGCTA. Below are three examples of paired statements and questions from the WGCTA practice test. These questions will be used to highlight the initial process of identifying how the questions were structured, and then how that information was used to develop questions for the DOT-CCTT. Each example from the WGCTA practice test and the DOT-CCTT questions which emerged from studying the WGCTA practice tests question will be discussed in turn. The three questions of the DOT-CCTT will be revisited at each stage of the study to highlight the changes made to the test based on the reliability and validity data collected throughout the study. With that in mind, it is important to note that the early versions of the questions of the DOT-CCTT may appear cumbersome and the underlying chemistry of these questions was continually corrected. Furthermore, the first version of the DOT-CCTT used essentially the same instructions as provided on the WGCTA practice test. In later versions of the DOT-CCTT the instructions were also changed as will be highlighted later in this chapter.

Below is the parent statement or stem of the 'Assumption' section of the WGCTA practice test. The introduction of this section provided an example of an assumption, stating that assumptions are presupposed facts. The instructions then advised that an assumption is made when it is logically justifiable and based on the evidence provided in the statement.

Monarch nations, i.e. those with royal families, differ from republican nations in several ways. An example of this difference is that citizens of monarchic nations pay more tax than citizens of republic nations.

This statement was then followed by the question asking if the assumption was made or not:

Republican nations do not have a royal family.

Assumption made Assumption not made

For this question the assumption was made. The explanation from the practice test suggesting as follows:

The statement says that monarchic nations are those with a royal family. The statement is assuming that this is one aspect which differentiates monarchic nations from republican nations.

This stem required the participant to recognise A and B (monarch nations and republican nations, respectively) are different, and explicitly states that there is relationship between A and C (monarch nations and royal families). The question asked if it is a reasonable assumption that B is not related to C. In writing an assumption question for the DOT-CCTT paramagnetic and diamagnetic behaviour of metal complexes replaced A and B. The relationship between A and C was replaced with paramagnetic behaviour being related to unpaired electrons. Furthermore, the pros were rearranged to improve readability of the stem. Below is the example of an assumption question for the DOT-CCTTv1:

A chemist tested a metal centred complex by placing it in a magnetic field. The complex was attracted to the magnetic field. From this result the chemist decided the complex had unpaired electrons and was therefore paramagnetic rather than diamagnetic.

This was then followed by the question:

Diamagnetic metals centred complexes do not have any unpaired electrons.

Valid Assumption Invalid Assumption

The correct answer was a valid assumption as this question required the participant to identify B and C were not related much like the example from the practice WGCTA. The explanation for the correct answer was as follows:

The paragraph suggests that if the complex has unpaired electrons it is paramagnetic. This means diamagnetic complexes likely cannot have unpaired electrons.

The next example is drawn from the 'Analysing Arguments' section of the WGCTA practice test. The instructions provided in this section highlight that a strong argument must be both important and related to the question being asked. The test taker is then told to decide if arguments are strong or weak based on the criteria previously described. Below is an example of a statement from WGCTA practice test:

Should companies downsize their workforces to decrease expenses and maximise their profits?

The following argument was then presented:

Yes, companies have no obligation to employ more people than it can handle.

Strong Argument Weak Argument

The practice test considered this a weak argument stating:

Although the point is taken to be true, it does not provide evidence of the benefits of its position, it merely states an irrelevant fact, making this a weak argument.

In writing a question for the DOT-CCTT the recent debate surrounding the toxicity of zinc oxide in sunscreens was considered. This argument was popular in the media after trade workers from BlueScope steel were wearing sunscreen and their hand prints appeared to leave rust marks where the sunscreen had made contact with metallic roofing. The aim behind writing this question was to maintain the argument of should D (companies) remove E (workforce) to decrease F (expenses). In the case of the DOT-CCTT question should zinc oxide (E) be removed from sunscreen (D) to decrease exposure to toxins (F)?

Zinc oxide (ZnO) protects against DNA damage from UV radiation. Should it be removed from all sunscreens to decrease exposure to toxins?

The argument presented in the DOT-CCTT was as follows:

Yes. There are other chemicals which could be used in sunscreens to absorb UV radiation before it can do damage to DNA.

Valid Argument Invalid Argument

Much like the question from the WGCTA practice test it was intended that the argument was invalid, and that the point raised may have been true but not directly related to the question being asked, citing the following explanation:

While there are other chemicals that could absorb UV radiation, the argument is whether removing zinc oxide will reduce exposure to toxins, which is not discussed.

The final example that will be discussed throughout this chapter is drawn from the 'Deduction' section of the WGCTA practice test. In this section of the test participants are asked to read a statement related to a stem and decide if the conclusion follows or does not follow. The participant is instructed to assume all statements as true and base their responses solely on the information provided. Furthermore, the participant is instructed to avoid using their own opinions or prejudices. Below is an example of the stem for this section:

Sarah owns a new company. New companies are more likely to fail than well-established companies. Therefore:

This was then followed by the statement:

Sarah's company is more likely to fail than a well-established company.

Conclusion Follows Conclusion Does Not Follow

In this question the conclusion follows and the WGCTA practice test provides the following explanation:

The statement notes that new companies are more likely to fail. As a new company, Sarah's company is therefore more likely to fail. The correct answer is therefore conclusion follows.

This question utilises the application of deductive logic. The stem states that A (Sarah's company) is B (a new company). It then states B is more strongly related to C (failure), than D (an established business) is related to C. The question then asks if A is more strongly related to C, than C is related to D. As A is B, and B is more strongly related to C than D, it holds that the conclusion follows. In writing the DOT-CCTT the 'Deduction' section was framed as testing hypotheses to highlight how scientists apply deductive reasoning. The following example is drawn from the esterification of an alcohol, and comparing the use of a carboxylic acid and an analogous anhydride. In this question A is replaced with aspirin, B is the formation of an ester bond using a carboxylic acid, C is likelihood of formation of an ester, and D is use of an anhydride:

When making aspirin a chemist needed to make an ester bond. However, she only had access to a carboxylic acid. The carboxylic acid is less likely to form an ester bond than if an anhydride were used.

The question on the DOT-CCTT then inquires as to the relationship between A and C:

The formation of aspirin using carboxylic acid is more likely to fail than if an anhydride were used.

Logical Deduction Illogical Deduction

For this section of the DOT-CCTT, the terms logical and illogical deduction were used to reflect the deductive thinking required in this section. In this particular question, the statement was a logical deduction for the following reason:

The passage clearly states that a carboxylic acid will be less successful than an anhydride.

All 85 question on the WGCTA practice test were analysed in the manner exemplified above. In writing these questions there were two motivations. Firstly, the test needed to be

able to be completed comfortably within 30 minutes to allow it to be administered in short time frames, such as at the end of laboratory sessions, and to increase the likelihood of completion by students. Secondly, the test needed to be able to accurately assess the critical thinking of chemistry students from any level of study, from first year general chemistry students to final year capstone research project students. To this end, the chemistry terminology needs to be written in such a way that a prior understanding of chemistry was not necessary to adequately comprehend the question and attempt the DOT-CCTT. The DOT-CCTT was written in such a way that any chemical phenomena were explained and contextualised completely within the stem and the questions. For example, the question previously discussed, which addresses assumptions around magnetism and electron pairing, only the information within the question was required and a chemistry expert in the area should not be advantaged any more than a novice chemistry student when attempting the question. The DOT-CCTT did assume a very basic understanding of science, in as much that chemical notations were provided alongside written descriptions of some chemicals, and appropriate scientific terminology was used, for example 'diamagnetic'. While chemical terminology was used, reading into its meaning was either not necessary to answer the question, or explicitly detailed to the participant.

Before administering the DOT-CCTTv1 to a group of students, members of the Monash Chemistry Education Research Group attempted the test in its entirety and their feedback was discussed with the researcher. This process was an informal discussion and considered an exploratory discussion of content validity. The group consisted for two teaching and research academics, one post-doctoral researcher and two PhD students. During these discussions the group received the intended responses to the DOT-CCTT questions. The participants shared which questions they felt did not illicit the intended critical thinking behaviour, citing poor wording and instances where the chemistry was poorly conveyed. Participants also expressed frustrations between the selection of five potential options in the 'Develop Hypotheses' section of the DOT-CCTT, having trouble distinguishing between options such as 'true' or 'probably true' and 'false' or 'probably false'. All participants identified that they took in excess of 40 minutes to complete the DOT-CCTT. It was a concern that individuals who could be

considered experienced critical thinkers required so much time to complete the test. Although it is known that the reliability and validity of a test is improved with a larger number of questions (Ennis, 1993; Pallant, 2016) the goal was to develop a test which could realistically be completed within 30 minutes, which is more reflective of the WGCTA-S. The WGCTA-S contains 40 multiple choice questions more suitable for being administered within 30 minutes (Pearson, 2015). The feedback provided by the Monash Chemistry Education Research Group was used to reduce the number of questions. Questions which were identified as unclear, which did not illicit the intended responses, or caused misconceptions of the scientific content were removed. Furthermore, several questions within the WGCTA practice test, and therefore the DOT-CCTT, tested similar critical thinking behaviours and were considered redundant. After refining the DOT-CCTT, it still retained the five sections or sub-scales, each with their own unique set of instructions, totalling 30 questions. 30 questions were chosen as opposed to the 40 questions used in the WGCTA-S to ensure the DOT-CCTT could be completed within 30 minutes. The CCTTv1 contained seven questions relating to 'Making Assumptions', seven questions relating to 'Analysing Arguments', six questions relating to 'Developing Hypotheses', five questions relating to 'Testing Hypotheses' and five questions relating to 'Drawing Conclusions'. Furthermore, the terms used to select a multiple choice option were written in a manner more accessible to science students, for example using terms such as 'Valid Assumption' or 'Invalid Assumption' instead of 'Assumption Made' or 'Assumption Not Made'. Finally, the number of options in the 'Developing Hypotheses' section were reduced from five options to three options of 'likely to be an accurate inference', 'insufficient information to determine accuracy' and 'unlikely to be an accurate inference'.

2.11 DOT-CCTTv1 Internal Reliability and Content Validity

First year chemistry students were chosen as the cohort on which to conduct initial internal reliability testing of the DOT-CCTTv1 (see Appendix E). Simultaneously, content validity studies were conducted with academic participants and education designers from a science education research community of practice discussion group at Monash University.

2.11.1 Undergraduate Student Participants

First year chemistry participants were drawn from semester one of 2016 in two units; Chemistry I and Advanced Chemistry I. Chemistry I was a general chemistry course with a mixed cohort of science students, most of whom completed chemistry studies in high school, and some who studied science but had not previously studied chemistry. Advanced Chemistry I consisted only of students who had completed chemistry in high school. Advanced Chemistry I covered the same content as Chemistry I with additional time in laboratory sessions and no tutorial.

2.11.2 Academic and Teaching Designer Participants

Academics and education designers from the Monash Science Teaching and Education Research community of practice were invited to a focus group held over two one-hour sessions. The data was considered representative of science academics with an interest in education and education researchers who were considered to possess expertise in critical thinking. However, the sample was highly self-selected as participation in the community of practice was an opt-in active, only academics with an inherent interest in teaching attended meetings. Furthermore data collection was convenient and opportunistic as the availability of some academics was limited with some unable to attend both sessions.

2.11.3 Internal Reliability Method

All first year participants were provided with the DOT-CCTTv1 at the conclusion of a compulsory laboratory safety induction session during orientation in the first week of semester one. Tests were administered to the entire first year cohort of approximately 1200 first year undergraduate chemistry students. Students were provided with the DOT-CCTTv1, an explanatory statement and completed the test on an optically read multiple choice answer sheet. 744 answer sheets were submitted. These sheets were digitally transcribed via optical reader to Microsoft Excel and the resultant data was imported into IBM SPSS statistics (v 22) as 744 cases. A frequency table was generated to determine erroneous or missing data. Nine erroneous data points were identified and removed from analysis as the participants had

selected 'C' to questions which only contained options 'A' or 'B'. 157 cases contained missing data. A variable was created to determine the sum of missing data for each case. Pallant (2016, pp. 58-59) suggests a judgement call is required when considering missing data and whether to treat certain cases as genuine attempts or not. Cases with one or two missing responses were considered genuine attempts as many participants only attempted between one and six questions. Cases with three or more missing responses were removed from further analysis. The removal of these cases resulted in 615 cases which were used for statistical analysis. As approximately half the first year cohort genuinely attempted the DOT-CCTTv1, it is proposed that this initial test was representative of the overall cohort (Krejcie and Morgan, 1970). This data was considered a representative random sample of first year chemistry students as the induction session was a prerequisite for all students commencing study in the first year chemistry laboratory. However, it must be acknowledged that the data may be reflective of self-selecting participants who could have been high achieving students or inherently interested in developing their critical thinking skills. At the time, demographic data such as sex, age, previous chemical/science knowledge, previous academic achievement and preferred language(s) were not collected and it is possible one or more of these discriminates may have impacted performance on the test. Whilst not addressed in the study of the DOT-CCTTv1, demographic data was subsequently collected and analysed for studies of the DOT-CCTTv2 and DOT-CCTTv3 and will be discussed later in this chapter.

Descriptive statistics were collected and used to determine whether the data was parametric or non-parametric. For the 615 genuine attempts of the DOT-CCTTv1 the overall scores of the test and the scores for each of the five sections were found to exhibit a normal (Gaussian) distribution (See Appendix F).

Internal consistency was determined as outlined in Section 2.3.1. The Cronbach's α of the DOT-CCTTv1 as single scale and the DOT-CCTTv1 made up of five sub-scales will be presented in Chapter 3. Changes in internal consistency as a result of deleting questions or sub-scales and CITC data for each question can be found in Appendix F, Table F4 and F5. An

item difficulty analysis was conducted in line with the procedure described in Section 2.3.2. Graphical representations and detailed analysis of this data can be found in Section 3.2.2.

2.11.4 Content Validity Method

Focus groups consisting of academic and teaching designer participants. These focus groups were conducted over two separate one hour meetings. The first meeting consisted of 15 participants and the second meeting consisted of nine participants. It is important to note that only five participants were able to attend both sessions. The remaining participants were unable to attend both sessions due to schedule conflicts. Participants were provided with the DOT-CCTTv1 and asked to complete the questions from a given section, for example 'Making Assumptions'. In the first session the 'Making Assumptions' and 'Analysing Arguments' sections of the DOT-CCTTv1 were discussed. In the second session the 'Testing Hypotheses' and 'Drawing Conclusions' sections were discussed. The 'Developing Hypotheses' section was not discussed in either session due to time constraints. After completing a section on the DOT-CCTTv1 participants were asked to discuss their responses, their reasoning for their responses and comment how they might improve the questions to better elicit the intended critical thinking response. The focus groups were recorded, transcribed and analysed in line with the procedures and theoretical frameworks described in Section 2.1 to result in four distinct themes which were used to code the transcripts (See Appendix G). The description of the themes and related excerpts will be presented in Section 3.2.3.

2.12 DOT-CCTTv2 Test-Retest Reliability, Convergent Validity and Content Validity

Several changes were made to the DOT-CCTTv1 upon analysis of the data obtained from students and academics to produce the DOT-CCTTv2. Firstly, the terms used to stimulate responses to various sections were refined in the instructions. For example, in the 'Making Assumptions' section what was considered a 'Valid Assumption' or an 'Invalid Assumption' was made more explicit through the use of examples relative to scientific practice. Responses such as 'Illogical Deduction' and 'Logical Deduction' were changed to 'Reasonable Deduction'

and 'Unreasonable Deduction' based on the feedback from academics that to use the terms logical and illogical were more suggestive of the application of syllogistic reasoning rather than deduction.

Several parent statements were rewritten to include additional information with the intention to reduce the need to draw on knowledge external to the questions. For example, in the questions that used formal charges on anions and cations, statements were included to describe the superscripts denoting charges: 'Carbonate (CO_3^{2-}) has a formal charge of negative 2.' The most extensive rewriting of the parent statements occurred in the 'Analysing Arguments' section. Initially these parent statements were more reflective of the WGTCA practice test, in that they were questions to prompt points of argument:

'Zinc oxide (ZnO_2) protects against DNA damage from UV radiation. Should it be removed from all sunscreens to decrease exposure to toxins?'

The feedback provided from the academic focus groups indicated that parent statements such as the one above did not include sufficient information for participants to adequately respond to the questions associated with the statement. In response to this feedback additional information was provided within the parent statements as can be seen below:

'Zinc oxide (ZnO) is the active ingredient in sunscreens which protects against DNA damage from UV radiation. Zinc oxide (ZnO) does this by effectively reflecting and diffracting harmful UV-radiation. Sunscreen residue left by tradespeople wearing sunscreen containing zinc oxide (ZnO) was thought to have caused rusting of many building materials. This has led lobbyists to be concerned that zinc oxide (ZnO) may in fact be toxic. Should it be removed from all sunscreens to decrease the risk of exposure to toxins?'

Additional qualifying statements were added to several questions in order to reduce ambiguity. For example, adding statements such as 'assume all other reaction conditions are the same' or 'assume that the voltage gated potassium ion channels of a typical cell membrane at rest are not open'. These statements aimed to restrict the need to draw on outside knowledge many of the focus group participants had a tendency to exhibit. In the stem

presented below from the 'Making Assumptions' section of the DOT-CCTTv2 the first sentence was added to eliminate the need to understand that differences exist between diamagnetic and paramagnetic metal complexes, overtly highlight that there is a difference with respect to how they interact with magnetic fields:

Paramagnetic and diamagnetic metal complexes behave differently when exposed to a magnetic field. A chemist tested a metal complex by placing it in a magnetic field. From the result of the test the chemist decided the metal complex had unpaired electrons and was therefore paramagnetic.

Similarly, the following question from the 'Testing Hypotheses' section of the DOT-CCTTv1 was rewritten:

The formation of aspirin using carboxylic acid is more likely to fail than if an anhydride were used.

The content validity analysis conducted on the DOT-CCTTv1 revealed that the term 'fail' appeared too ambiguous to experienced scientists. Therefore the question was rewritten to focus on the success of using an anhydride:

Anhydrides are more likely to succeed in forming an ester bond than carboxylic acids.

Finally, great effort was made in the organisation of the DOT-CCTTv2 to guide the test taker through a critical thinking process. Similar to Halpern's approach to analysing an argument (Halpern, 1996a), the test taker was provided written scaffolding from making assumptions to analysing arguments. Each sections of the DOT-CCTTv2 was made relevant to scientific practice within the instructions.

2.12.1 Undergraduate Student Participants

Participants for the study of the DOT-CCTTv2 were recruited by means of advertisements in Monash chemistry facilities and learning management system pages. The invitation was open to any current Monash student who was either currently studying a

chemistry unit or had previously completed a chemistry unit at Monash University. Participation was voluntary and it was clear to participants that it would in no way affect their academic record. 40 students expressed their interest in attending two focus groups on consecutive days and provided their contact details via a Google Form. Of these 40 students 20 of them attended the first day of the study and 18 of these attended the second day. All activities and recruitment processes were approved by MUREHC (Section 2.6).

2.12.2 Structure of the Two Day Study

Day one started with an introduction and students were asked to complete consent forms and provide demographic data: age, sex, dominant language, previous academic achievement using ATAR scores (Australian Tertiary Admission Rank) or equivalent, current level of chemistry being studied and highest level of chemistry study completed at Monash University. They then completed the DOT-CCTTv2 using an optical reader multiple choice answer sheet. This was then followed by completion of the WGCTA-S in line with procedures outlined by the Watson-Glaser critical thinking appraisal short form manual (2006). The WGCTA-S was chosen for analysis of convergent validity, as it was similar in length to the DOT-CCTTv2 and was intended to measure the same aspects of critical thinking. After a brief break, the participants were divided into groups of five to eight students and interviewed about the WGCTA-S for approximately 45 minutes. They were asked about their overall impression of the test and their approach to various questions. At this time interviewers prevented the participants from discussing the DOT-CCTTv2 so as not to influence participants' responses upon retesting. In total the first day required four hours of the students' time. Day two consisted of participants repeating the DOT-CCTTv2. DOT-CCTTv2 attempts were completed on consecutive days to minimise participant attrition. Upon completion of the DOT-CCTTv2 and after a short break participants were divided into two groups of nine and interviewed. This interview focused on the DOT-CCTTv2, the participants' thoughts on the test, how they approached various questions and comparisons between the DOT-CCTTv2 and WGCTA-S. Day two required an hour and half of student's time.

2.12.3 Test-Retest Reliability and Convergent Validity Method

Responses to the tests and demographic data were either entered directly into Microsoft Excel or transcribed from an optical reader file then imported to IBM SPSS statistics (v22). Frequency tables were generated and descriptive statistics were determined for responses. There was no missing data with respect to responses to test questions (with the exception of those participants who did not repeat the DOT-CCTTv2). With the exception of ATAR score or equivalent, there was no missing or erroneous demographic data. Two participants did not disclose their ATAR score and three participants identified that they did not complete their studies in Australia.

Spearman's ρ correlations were performed as described in Section 2.4.3 comparing ATAR scores to scores on the WGCTA-S and the DOT-CCTTv2. Although the sample size was too small to make any generalisation with respect to whether previous academic achievement was related to performance on either test, it was of interest to explore the discriminate validity at this early stage of test development.

Test-retest reliability was determined using a Wilcoxon signed rank test, in line with the rationale outlined in Section 2.3.3. Median scores of the participants' first attempt of the DOT-CCTTv2 on day one were compared with the median score of the participants' second attempt of the DOT-CCTTv2 on day two. To determine the convergent validity of the DOT-CCTTv2 the relationship between performance on the DOT-CCTTv2 and performance on the WGCTA-S was investigated using Spearman's ρ correlation as outlined in Section 2.4.1.

2.12.4 Content Validity Method

In each of the interviews participants were provided with blank copies of the relevant test (the WGCTA-S on day one and the DOT-CCTTv2 on day two). Participants were advised they were being recorded, that their comments in no way affected their academic record, and that they would be de-identified during analysis in accordance with the ethics presented in Section 2.6. Participants were encouraged to make any general remarks or comments with respect to the tests they had taken, with the exception of the interviews on day one, where interviewers prevented any discussion of the questions with the DOT-CCTTv2. It is important

to note that beyond the guidelines described in this section, the interviewers did not influence the participants' discussion, correct their reasoning or provide the correct answers to either test.

After approximately 15 minutes of participants freely discussing the relevant test, the interviewers asked the participants to look at given section on a test, for example the 'Testing Hypotheses' section of the DOT-CCTTv2, and identify any questions they found problematic. In the absence of students selecting any questions, the interviewers were provided with a list of questions from both the DOT-CCTTv2 and the WGCTA-S to prompt discussion. Once a question had been selected to discuss, the interviewers asked the participants to indicate their response to the question by show of hands. The participants were then asked as a group:

- 'What do you think the question is asking you?'
- 'What do you think is the important information in this question?'
- 'Why did you give the answer(s) you did to this question?'

The interviews were transcribed and analysed in line with the procedures and theoretical frameworks described in Section 2.12.4.4 to result in four distinct themes which were used to code the transcripts (See Appendix J). The description of the themes and related excerpts will be presented in Section 3.3.4.

2.13 DOT-CCTTv3 Internal Reliability, Criterion Validity and Discriminate Validity

The DOT-CCTTv2 was edited to reflect the feedback from the student focus groups, with particular attention to the instruction sections and the use of scientific terminology. In the DOT-CCTTv3 a cover page was added with instructions for the overall test. These instructions introduced the significance of critical thinking in scientific practice and outlined the structure of the test. As it was found that student participants drew heavily on the worked examples in the introduction of each section in both the DOT-CCTTv2 and the WGCTA-S, carefully written examples were provided for each section of the DOT-CCTTv3. These explanations highlighted the underlying reasoning behind a choice and provided context relevant to science. For

example in the 'Testing Hypotheses' section an 'Unreasonable Deduction' regarding the identity of a product using thin layer chromatography was supported by recognition that there 'is insufficient evidence to support this claim.'

The other major change was the simplification or removal of certain scientific terminology. Some questions were completely rewritten to reduce the temptation for the test taker to use external information. For example questions 19 and 20 removed the use of a 'carbon-centred molecule' and instead focused on molecules, activation energy and catalysts more generally. The question still required the identification of the limitations of correlations as in DOT-CCTTv2 but uses simplified terminology. Likewise, questions 1 to 4 focused on an alloy of thallium and lead rather than a 'metal complex'. Generalising questions 1 to 4 to focus on an alloy allowed these questions to retain scientific accuracy and reduce the tendency for participants to draw on knowledge outside the information presented in the questions:

Metals which are paramagnetic or diamagnetic behave differently when exposed to an induced magnetic field. A chemist tested a metallic alloy sample containing thallium and lead by placing it in an induced magnetic field. From the test results the chemist decided the metallic alloy sample repelled the induced magnetic field and therefore was diamagnetic.

This statement was followed by the prompt asking the participant to decide if the assumption presented was valid or invalid:

Paramagnetic metals do not repel induced magnetic fields.

The essence of this question had not changed from the DOT-CCTTv1 and the assumption is valid. The explanation for the response was as follows:

The paragraph states that the alloy was deemed diamagnetic as it repelled the induced magnetic field. The paragraph states that that diamagnetic and paramagnetic metals behave differently therefore it is reasonable to assume that paramagnetic metals do not repel induced magnetic fields.

Similarly, the stem of questions 14 to 16 of the 'Testing Hypotheses' section were completely written:

A chemist needed to make aspirin using a carboxylic acid. One of the by-products of this chemical reaction is water. The carboxylic acid is less likely to form aspirin when the water by-product is not removed during the chemical reaction.

The stem still essentially is discussing the formation of an ester, however the term 'ester' has been removed from the stem. Instead of discussing the likelihood of reaction comparing the use of a carboxylic acid and an anhydride, the stem now considers the role of water in the equilibrium of this reaction. From the above stem the test taker is asked to determine if the following statement is reasonable or unreasonable deduction:

Removing water during the chemical reaction is more likely to succeed in forming aspirin than when water is not removed during the chemical reaction.

This question is considered a reasonable deduction as the underlying logic of the question has not changed:

The passage clearly states that the carboxylic acid will be less likely to form aspirin when water is not removed.

Several terms were rewritten as their use in science implied several assumptions as identified by the student focus groups. These assumptions were not intended to be part of these questions and hence were reworded. An example is question 14 which asked whether a 'low yield' would occur in a given synthetic route. The term 'low yield' was changed to 'an insignificant amount' to remove any assumptions regarding the term 'yield'. In question 24 regarding the use of nano particles, the phrase 'small enough to be absorbed by skin cells' was replaced with 'behave differently at such a small scale' as there were immediately negative assumptions associated with chemicals absorbed by skin cells.

The study of the DOT-CCTTv3 required participants to be drawn from several distinct groups in order to assess criterion and discriminate validity. For the purpose of criterion validity, the DOT-CCTTv3 was administered to first year and third year undergraduate chemistry

students, honours and PhD students and post-doctoral researchers at Monash University, and chemistry education academics from an online community of practice. Furthermore, third year undergraduate chemistry students from Curtin University also completed the DOT-CCTTv3 to assist in determine discriminate validity with respect to performance of the DOT-CCTTv3 outside of Monash University.

2.13.1 Undergraduate Student Participants

First year participants were drawn from the Chemistry I unit at Monash University. As described in Section 2.11.1, Chemistry I was a general chemistry unit run in semester one and in 2017 had 1059 students enrolled in the unit, 57% ($n = 603$) female and 43% ($n = 456$) male. The DOT-CCTTv3 was administered to the students at the end of a laboratory session. The test was presented to 576 students in a paper-based format. Of these students 199 (35%) attempted the test, representing approximately 19% of the entire first year cohort. The data obtained was representative of the distribution of sexes in the first year cohort as 54% of respondents identified as female, 43% identified as male and 2% did not provide their sex. Completion of the DOT-CCTTv3 was untimed and students typically took between 20 and 30 minutes to complete the test.

Third year participants were drawn from an advanced inorganic chemistry course at Monash University and a capstone chemical engineering course at Curtin University. 146 students undertook Advanced Inorganic Chemistry at Monash University and had completed second year prerequisite units in chemical synthesis and analysis. 42% ($n = 62$) of the students identified as female and 58% ($n = 84$) identified as male. The DOT-CCTTv3 was administered at the conclusion of a one hour lecture towards the end of semester one in 2017 for a period of 30 minutes. The test was administered in a paper-based format and the students responded to the multiple choice questions directly onto the test. A total of 54 (37%) third year students attempted the DOT-CCTTv3. The data obtained was representative of the distribution of sexes in the third year cohort as 39% of respondents identified as female and 61% identified as male.

The 23 students who completed the DOT-CCTTv3 represented the entire cohort of students who undertook Chemistry Research Methods offered at Curtin University. Students

comprised of third year chemistry and biochemistry majors, and fourth year chemistry/chemical engineering majors. Pre-requisites to undertake the unit required that students had successfully completed a third year synthetic methods, or chemical sensing and measurement course, and a third year analytical chemistry and spectroscopy course. Students in this unit undertook a research project under the guidance of an academic for approximately 15 hours per week and were required to attend a one hour lecture and a one hour workshop each week. The DOT-CCTTv3 was administered in one of the workshops towards the start of semester two in 2017 for a period of 30 minutes. The DOT-CCTTv3 was administered in a paper-based format and the students responded to the multiple choice questions directly onto the test. All students completed the DOT-CCTTv3 in full.

2.13.2 Post-Doctoral Researchers, Honours and PhD Student Participants

Post-doctoral researchers, honours and PhD students from Monash University were invited to attempt the DOT-CCTTv3. Students who undertook honours were self-selected from those who completed third year chemistry with a distinction (70 %+) grade point average. In 2017 there were 25 students who undertook honours. Admittance to the PhD program in the School of Chemistry is competitive and PhD students have obtained at least a high distinction 3.8 grade point average. Finally, post-doctoral researchers in the School of Chemistry must have a completed PhD. The length of incumbent and specific duties of these post-doctoral researchers was highly variable. Key among all of these groups was that their primary role was to conduct research and communicate that research through a thesis and/or peer reviewed publications. Post-doctoral researchers, honours and PhD students were invited by email to attend a session to undertake the DOT-CCTTv3. These sessions were held during lunch where refreshments were provided. 40 participants drawn from these cohorts attended one of these sessions and completed the DOT-CCTTv3 in a paper-based format, marking responses directly onto the test. Participants required approximately 20 to 30 minutes to complete the test.

2.13.3 Academic Participants

An online academic cohort consisted of around 300 members of a chemistry education email discussion group predominately from Australia, the UK and Europe. These participants received a link to an online version of the DOT-CCTTv3 sent via a third party. Online completion was untimed and 46 participants completed the online DOT-CCTTv3.

2.13.4 Treatment of Data

All responses to the DOT-CCTTv3 were transcribed into Microsoft Excel and the resultant data was imported into IBM SPSS statistics (v 22) as 363 cases. A frequency table was generated to determine erroneous or missing data. Two erroneous data points were identified which were questionnaires conducted with undergraduate students who identified their education/occupation as that of an academic. It was likely these undergraduates either mistakenly identified as academics or did not take the DOT-CCTTv3 seriously. Consequently, these two cases were removed. 97 cases contained missing data. A variable was created to determine the sum of missing data. Cases with seven or fewer missing responses were considered genuine attempts as many participants only attempted between one and six questions. Seven missing responses was considered a genuine attempt in this instance as many Monash third year participants did not complete the DOT-CCTTv3 due to time restrictions within the lecture setting. A total of 67 cases were not considered genuine attempts of the test due to missing data and were removed from any further statistical analysis, as recommended in the literature (Pallant, 2016, pp. 58-59). This resulted in 296 cases which were used for statistical analysis.

Participants were provided with several options to categorise their level of education or employment, as can be seen in Table 2. 1. These variables were recoded into the variables defined as 'Education Group'. As will be discussed in detail in Section 3.4.5, there was no statistical difference between the performance of third year students from Monash and Curtin Universities, and therefore the third years were treated as one group. Participants identifying as second year students were excluded from any further statistical analyse as they were drawn

from both first year and third year courses, and to alleviate the possibility they had attempted the DOT-CCTTv1 in 2016.

The Honours, PhD and Post-Doctoral variables were combined into the education group 'Postgraduates' as the data sets were so small. Six participants who were invited as part of the Honours, PhD and Post-Doctoral cohorts identified as teaching associates and these were also added to the education group 'Postgraduates'. Four participants invited via the online community of practice identified their role as 'Other'. These four cases were combine to form the education group 'Academics'. It is important to note that any generalisation made with respect to the 'Postgraduates' group were required to be conservation, as the honours, PhD and post-doctoral participants were so varied.

Table 2. 1 Summary of participants' level of tertiary education/employment and the resulting 'Education group' used for statistical analysis

Education	<i>n</i>	Education Group	<i>n</i>
First year	119	First year	119
Second year	18		
Third year (Monash)	44	Third year	67
Third year (Curtin)	23		
Honours	14	Post graduates	44
PhD	19		
Post-Doctoral	5		
Teaching Associate	6		
Academic	36	Academics	40
Other (Academic)	4		

Descriptive statistics were collected and used to determine if the data was parametric or non-parametric. The tests of normality (see Appendix L Table L1 and Figure L1) revealed the overall scores of the DOT-CCTTv3 exhibited a non-normal (non-Gaussian) distribution. The scores were skewed greater than the mean (skewedness = -.17) indicating that the scores of the DOT-CCTTv3 were not equally distributed either side of the mean, and a larger proportion of the scores were greater than the mean. When looking at the distribution of scores according to education groups, similar negatively skewed (-.04 to -.68) data was obtained for each group. One of the key indicators of 'normal' or parametric data is that data be distributed equally either side of the mean. As a larger proportion of the DOT-CCTTv3 scores were above the mean, the data had to be considered 'not normal' or non-parametric for the purposes of

reliability and validity statistical analysis. Furthermore this finding supported the choice to view all data as non-parametric made earlier in this research (Section 2.5).

2.13.5 Internal Reliability Method

The internal reliability was determined by calculating the internal consistency (Cronbach's α) and corrected item total correlations (CITC) for each section of the DOT-CCTTv3 similar to the study of DOT-CCTTv1 (Section 2.11.3). The internal reliability of the DOT-CCTTv3 was further explored by calculating the CITC and determining changes in α if questions were deleted for each sub-scale of the test. All internal reliability analyses were in accordance with the method outlined in Section 2.3.1. Furthermore, item difficulty analysis was performed using the effect size of each question using Mann-Whitney U tests and the percentage of participants who answered questions correctly in accordance with the approach described in Section 2.3.2.

2.13.6 Criterion Validity Method

Several Mann-Whitney U tests were conducted to determine the criterion validity of the DOT-CCTTv3 following the procedure in Section 2.4.2. The assumption was that academics were better critical thinkers than postgraduates; that postgraduates were better critical thinkers than third year undergraduates, and third year undergraduates were better critical thinkers than first year undergraduates. Based on this assumption, the hypothesis was that there would be a statically significant improvement in DOT-CCTTv3 scores relative to years of tertiary study.

To further explore the criterion validity education groups were compared based on the percentage of questions answered correctly within a given section of the DOT-CCTTv3. The percentage of participants within a given education groups who obtained the correct response for each question of the DOT-CCTTv3 was determined and tabulated. This data was used to determine if there were any trends with respect to the percentage of participants responding correctly. It was hypothesised that the percentage of correct responses should increase with respect to level of education group: first year undergraduates obtaining lowest percentage of

correct responses to a question or section, and academics obtaining the highest percentage of correct response to a question or section.

2.13.7 Discriminate Validity Method

Discriminate validity of the DOT-CCTTv3 was based on whether the achievement on the DOT-CCTTv3 was independent of age, previous academic achievement, and which university the participant attended. To determine the effect of age on performance on the DOT-CCTTv3 a Spearman's Rank-order correlation between score obtained on the DOT-CCTTv3 and age was performed ($\rho = .50$, $n = 284$, $p < .01$). As there was such a large positive correlation between age and score on the DOT-CCTTv3, further Spearman's Rank-order correlations between age and score obtained on the DOT-CCTTv3 were performed, looking at participants 30 years of age or less. The age restriction was made as 75% of participants were 30 years of age or less. The Spearman's Rank-order correlation coefficient found a moderate positive correlation between age and score on the DOT-CCTTv3 for participants 30 years of age or less ($\rho = .43$, $n = 220$, $p < .01$). As this correlation was still very strong, the age was further restricted to 25 years of age or less, which was reflective of ages typical of the undergraduate students. The Spearman's Rank-order correlation coefficient still found a moderate positive correlation between age and score on the DOT-CCTTv3 for participants of 25 years of age or less ($\rho = .37$, $n = 199$, $p < .01$). The effect of previous academics achievement, as measured by ATAR score, was determined via a Spearman's Rank-order correlation between score obtained on the DOT-CCTTv3 and ATAR score. Finally, the effect of the higher education institution which the participant attended was considered using a Mann-Whitney U test comparing the difference in the median score obtained on the DOT-CCTTv3 of 3rd year Monash University chemistry students and the median score obtained on the DOT-CCTTv3 of 3rd year Curtin University chemistry students.

2.14 Chapter Summary

Within this chapter the methodology was presented that was used to design a valid and reliable test that could be used to measure undergraduate chemistry students' critical thinking

skills independent of extensive chemistry knowledge, while still set within a chemistry context. The research followed five stages using a combination of qualitative and quantitative methods. Qualitative data was gathered using open-ended questionnaires to assist in the understanding the definitions of critical thinking. Statistical analyses were used in determining the reliability and validity of the various iterations of the DOT-CCTT and content validity was determined qualitatively through focus groups. Changes were made to the DOT-CCTT based on the feedback and data obtained from these analyses. The next chapter presents the results and analyses of this data.

Chapter 3 Results and Discussion

This chapter describes the results and discussion of the reliability and validity studies of each iteration of the Danczak-Overton-Thompson chemistry critical thinking test (DOT-CCTT). The discussion will highlight what changes needed to be made to the DOT-CCTT at each stage of development, the limitations of each analysis, and the significance of the results in the context of the literature. This chapter will be concluded with a discussion of the implications the DOT-CCTT has for teaching and education research in chemistry.

3.1 Introduction

This research project aimed to design a valid and reliable test to measure undergraduate chemistry students' critical thinking skills independent of extensive chemistry knowledge, while still set within a broad chemistry context. The goal was that the test could be administered to undergraduate chemistry students at any point during their studies to provide higher education practitioners with information regarding the development of their students' critical thinking skills, and therefore, develop teaching interventions to address areas of improvement identified by the DOT-CCTT. After researching the commercially available critical thinking tests, the Watson Glaser Critical Thinking Appraisal (WGCTA) was used as a template for the development of the DOT-CCTT, as detailed in Section 2.10. The DOT-CCTT comprised of 30 multiple choice questions, divided into five sections: 'Making Assumptions', 'Developing Hypotheses', 'Testing Hypotheses', 'Drawing Conclusion' and 'Assessment of Argument'.

As the DOT-CCTT was considered to be a psychometric test, validity and reliability testing were essential to ensure the DOT-CCTT accurately and precisely measured critical thinking of a student. Three separate reliability and validity studies were conducted throughout the course of this research thesis, resulting in three versions of the DOT-CCTT being produced. The first version of the DOT-CCTT (DOT-CCTTv1) underwent internal reliability studies with the first year chemistry students at Monash University, in tandem with content validity studies with a group of education focused academics in a Monash University community of practice. Based on the findings from these studies the second version of the

DOT-CCTT (DOT-CCTTv2) was developed and test-retest reliability, convergent validity and content validity studies were conducted with a focus group of a cross section of Monash University undergraduate chemistry students. The data from the study of the DOT-CCTTv2 was used to assist in developing the third version of the DOT-CCTT (DOT-CCTTv3) which was administered to first year, third year and post graduate chemistry students at Monash University, a third year chemistry student cohort at Curtin University and an online education academic community of practice. The data obtained from these groups were used to determine the internal reliability, criterion validity and discriminate validity of the DOT-CCTTv3.

3.2 DOT-CCTTv1

As recommended by Ennis (1993) the DOT-CCTTv1 underwent internal reliability and content validity studies with a group of the intended respondents and was scrutinised by academic experts in education and/or critical thinking. Internal reliability data was generated from DOT-CCTTv1 tests completed by first year undergraduate chemistry students at Monash University. Alongside the reliability study, content validity data was gathered from qualitative studies from focus groups of Monash University education-focused academics and an education designer.

3.2.1 Internal Reliability

As can be seen from Table 3. 1, when the internal consistency was determined the Cronbach's α of the DOT-CCTTv1 as single scale and the DOT-CCTTv1 made up of sub-scales were .63 and .68, respectively. These α values suggested the DOT-CCTTv1 viewed either as a single scale or five sub-scales had limited internal reliability. Therefore, questions or the sections of the DOT-CCTTv1 could not confidently be added together to measure critical thinking skill. The internal consistency as a result of deleting questions or sections and CITC data for each question can be found in Appendix F, Table F4 and F5. From the α values of each sub-scale presented in Table 3. 1, it would appear that the questions within each sub-scale did not reliably add together to measure the sub-scale of interest. However, as described previously, internal consistency values lower than $\alpha = .7$ can arise when a scale or sub-scale

contain less than ten questions. Each sub-scale of the DOT-CCTTv1 contained between five to seven questions, thus CITCs were also calculated for each sub-scale. Table 3. 1 shows that on average the questions of the 'Making Assumptions', 'Developing Hypotheses' and 'Testing Hypotheses' sections appear to be contributing to the measurement of their respective sub-scales as the mean CITC is greater than .3. Although, it is important to note that some questions within each section were not likely to be contributing to the measurement as some CITC values are as low as .21. The mean CITC of the 'Drawing Conclusions' and 'Analysing Arguments' sections suggested that on average the questions did not reliably contribute to the measure of the sub-scale of interest. In summary, the internal consistency and CITC data suggested that when considering the DOT-CCTTv1 to either be made up of one scale or five sub-scales, the DOT-CCTTv1 exhibited limited internal reliability. Split-halves analysis showed the DOT-CCTTv1 was found to have a value of $\rho_{cc'} = .47$. As the value was below the recommended value of $\rho_{cc'} = .8$ it further suggested that the DOT-CCTTv1 had low internal reliability.

Table 3. 1 Summary of internal reliability data for the DOT-CCTTv1

	Sub-scales	Internal Consistency (Cronbach's α)	Mean CITC	Min CITC	Max CITC
DOT-CCTTv1 (Single scale)		0.63			
DOT-CCTTv1 (Five sub-scales)		0.68			
	Making Assumptions (Q1-7)	0.65	0.30	0.21	0.40
	Analysing Arguments (Q8-14)	0.60	0.24	0.15	0.32
	Developing Hypotheses (Q15-20)	0.68	0.36	0.23	0.47
	Testing Hypotheses (Q21-25)	0.66	0.33	0.22	0.41
	Drawing Conclusions (Q25-30)	0.63	0.28	0.21	0.44

3.2.2 Item Difficulty Analysis

Figure 3. 1 shows the percentage of participants who answered each question correctly on the DOT-CCTTv1. Questions 1, 6, 9, 11, 18, 20, 21 and 30 fell below the 50% probability

of the questions been answered correctly by chance. When considering item difficulty analysis as described by Loo and Thorpe (1999), these eight questions were considered too difficult or poorly constructed and, therefore, not contributing to measurement of the psychometric construct of interest: critical thinking skill. The questions identified through this analysis received the most scrutiny in revising the DOT-CCTTv1 in preparations for the DOT-CCTTv2.

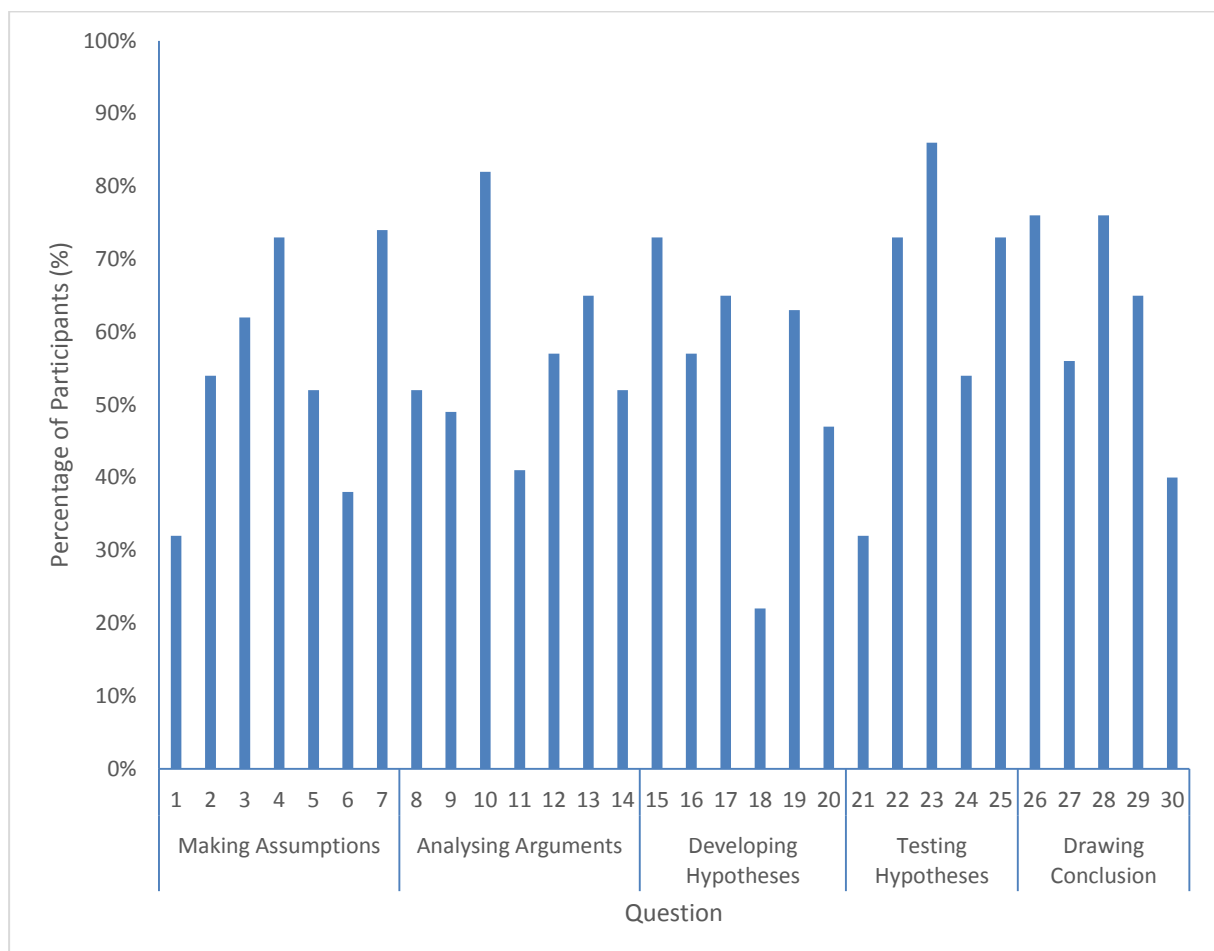


Figure 3. 1 Percentage of students who answered questions correct on DOT-CCTTv1

Figure 3. 2 shows the median score of participants who answered each DOT-CCTTv1 question correctly and the corresponding median score of participants who answered the same question incorrectly. The degree to which each question was able predict scores on the DOT-CCTTv1 was also explored using Mann-Whitney *U* tests. By treating each question as an independent variable which effects a dependent variable, in this case the score on the DOT-CCTTv1, the strength of the effect was determined from the effect size obtained from Mann-Whitney *U* test data (Cohen, 1988). For example, question 5 was found to have an effect size

of .53 when comparing the median DOT-CCTTv1 scores of participants who answered question 5 correctly against the median DOT-CCTTv1 scores of participants who answered question 5 incorrectly using a Mann-Whitney U test. This suggested that answering question 5 correctly was a strong predictor of a participant being part of the group who obtained a higher median score on the DOT-CCTTv1.

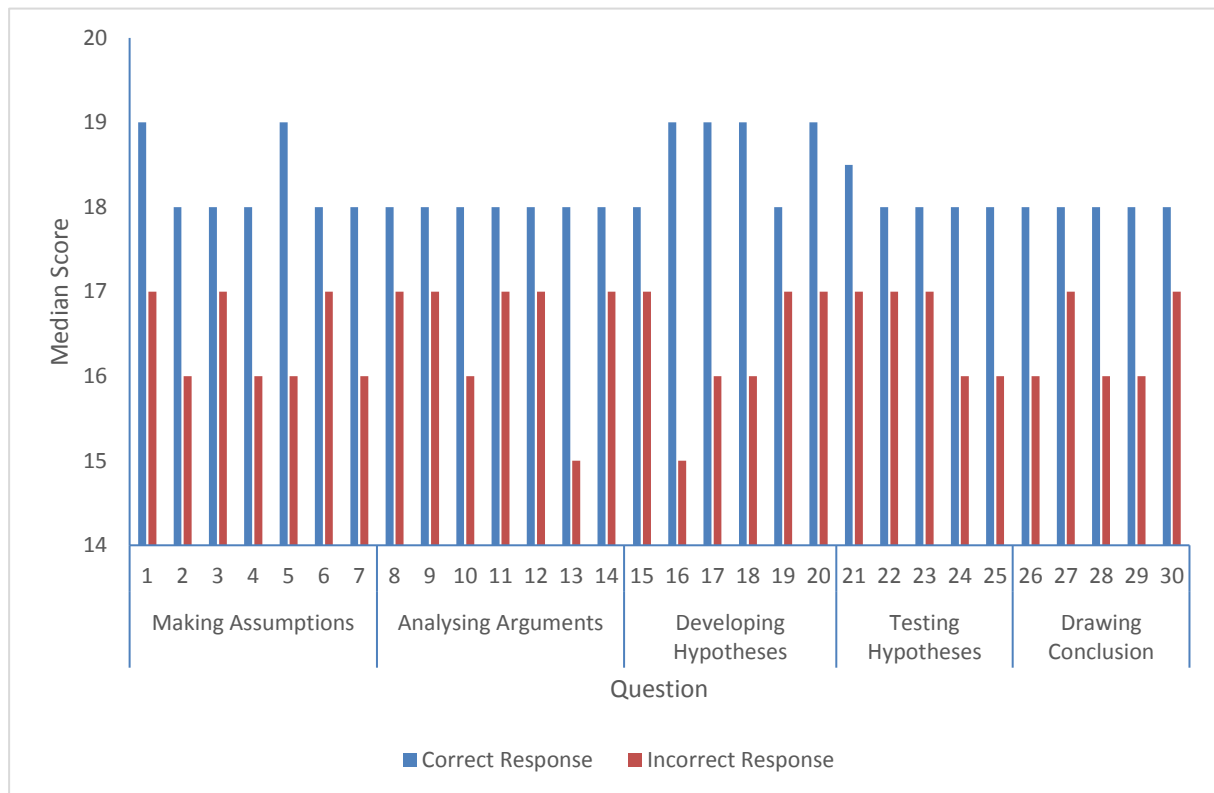


Figure 3. 2 Median score of participants who answered each DOT-CCTTv1 correctly versus incorrectly

Mann-Whitney U tests revealed all test questions to have a significant difference in the median score of participants who answered each DOT-CCTTv1 question correctly versus the median score of participants who answered the corresponding question incorrectly ($p < .001$). As can be seen in Figure 3. 3, questions 5 and 6 had large effect sizes ($r > .5$) (Cohen, 1988), suggesting that participants who answered questions 5 and/or 6 correctly were more likely to have obtained a higher score on the DOT-CCTTv1. Questions 3, 4, 7, 13, 15, 16, 17, 18 and 28 were found to have a medium effect size ($r > .3$). The medium effect size suggesting a moderate relationship between obtaining a higher score on the DOT-CCTTv1 and answering these question correctly. The remaining questions had a small effect size ($r > .1$), indicating

that answering these questions correctly was not a strong predictor of obtaining high scores on the DOT-CCTTv1.

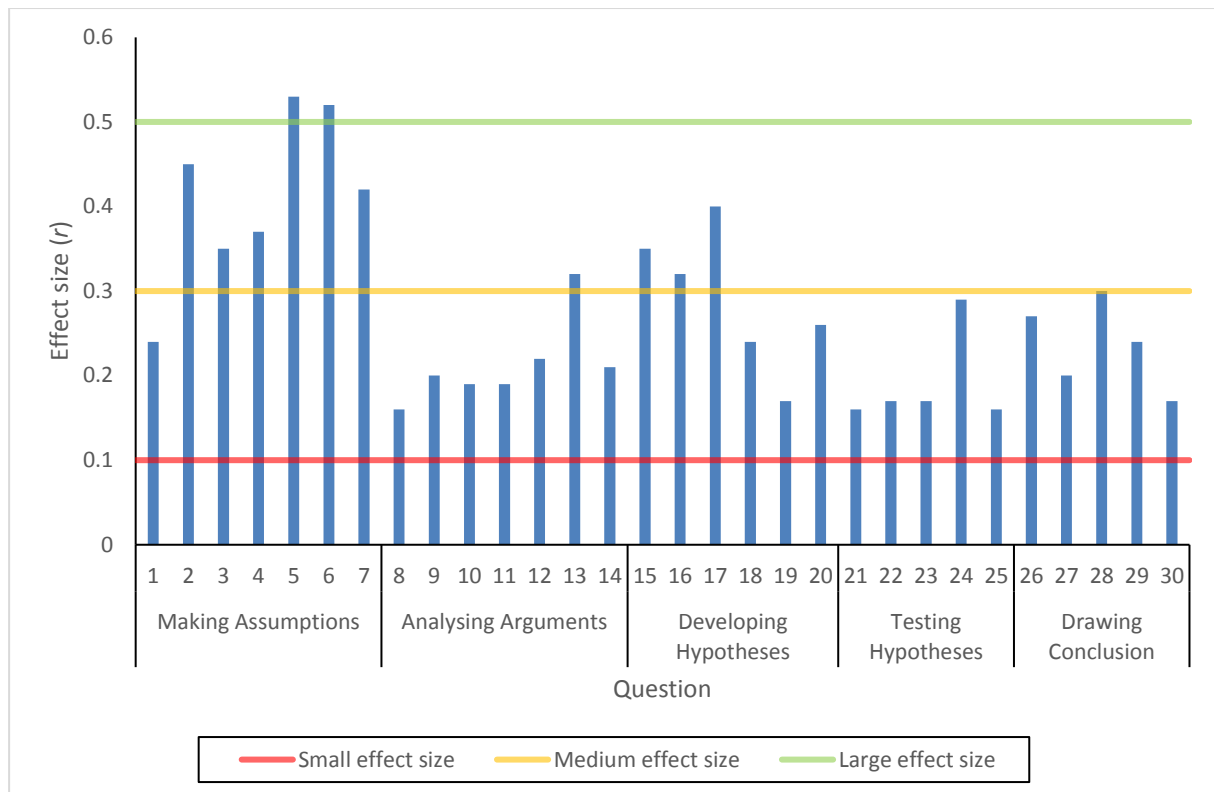


Figure 3. 3 Effect sizes (r) of the difference between median scores of participants who answered each DOT-CCTTv1 question correct or incorrect

Drawing any conclusions must be done conservatively, due to the small number of questions within each section of the DOT-CCTTv1 (Pallant, 2016, p. 104). However, the small effect sizes highlighted that 19 questions from 'Testing Hypotheses', 'Drawing Conclusions' and 'Analysing Arguments' sections required revisions as these questions were poor predictors of obtaining a higher score on the DOT-CCTTv1. Typically researchers who have used the WGCTA to evaluate their participants' critical thinking found that the 'Inference' section on the WGCTA (which is akin to the 'Developing Hypotheses' section on the DOT-CCTTv1) was the smallest contributor to overall critical thinking score, as measured by the section's effect size (Bernard *et al.*, 2008; Frye, Alfred and Campbell, 1999; Loo and Thorpe, 1999; Macpherson and Owen, 2010). The statistically small contributions observed by these researchers was often explained as the 'Inference' section being more challenging due to five multiple choice

options as opposed to dichotomous questions. The findings of the data obtained from the DOT-CCTTv1 appeared to run contrary to the majority of the literature with the exception of one author (Hassan and Madhum, 2007), who used the WGCTA for pre- and post-testing on a sample of 273 business students. The researcher reported an inter-item correlation of .63 for the 'Inference' section of the WGCTA, relative to other sections on the test and overall performance on the WGCTA. However, no explanation for this observation was made.

With respect to the 'Making Assumptions' section, Halpern's definition of an assumption is worth considering; "a statement for which no evidence or proof is offered (Halpern, 1996b, p. 172)". Halpern further comments that assumptions are often implied within a statement. It is possible that participants in the DOT-CCTTv1 did not determine the quality of the assumptions based only on the information given in the statements. As proposed by one of the participants in the content validity study of DOT-CCTTv2: *"Is it really an assumption if it is a true fact?"* This participant's query indicated they may have had difficulty when asked to evaluate the validity of an assumption that they may have considered scientific fact. What is clear from the study of DOT-CCTTv1 is that it would have been useful to ask students to provide contact details to enable follow up with individual interviews or focus groups in order to probe the approaches these students took. In retrospect, discussing the DOT-CCTTv1 with students would have aided in determining the content validity. Fortunately, possible explanations as to why certain questions were more difficult or poorly constructed, and what may have contributed to the effect sizes of various questions were provided from the content validity study conducted with the academic focus groups.

3.2.3 Content Validity

The academic focus group highlighted the issues with many of the DOT-CCTTv1 questions identified as problematic via statistical analysis of the first year students' performance. Themes identified from the academics' discussion focused on the clarity of instructions and relevant examples, the use of assumed science and/or chemistry terminology, clear connections between statements and questions, and use of binary terms such as 'rather than'.

As a group the academic participants navigated the questions on the DOT-CCTTv1 to arrive at the intended responses. However, there was rarely consensus within the group and a minority, usually one or two participants, disagreed with the group. The difficulties the academics had in responding to the DOT-CCTTv1 were made clear from four themes which emerged from the analysis: 'Instruction Clarity', 'Wording of the Question(s)', 'Information within the Statement' and 'Prior Knowledge'. The last theme was generally found to be associated with the other themes. A summary of the themes, a brief descriptor and a quote from the focus groups can be found in Table 3. 2.

Table 3. 2 Themes identified in the qualitative analysis of the academic focus groups for DOT-CCTTv1

Theme	Description	Example
Instruction Clarity	Insufficient information within the instructions (sometimes due to superficial reading)	<i>"...what do you mean by 'is of significant importance'?"</i> <i>"...when I read it properly I thought 'yes of course'."</i>
Wording of the Question	Attributing meaning to specific words within a question	<i>"Complex is the key word here right?"</i> <i>"...the language gives it away cause you say 'rather than'."</i>
Information within the Parent Statement	Adequate or insufficient information in the parent statement to response to the questions	<i>"Basically what you're providing in the preamble is the definition, right?"</i> <i>"...none of the answers should matter because you haven't given any context."</i>
Prior Knowledge	The using or requiring use of prior scientific knowledge from outside the test	<i>"Because actually all paramagnetic substances have a diamagnetic contribution."</i> <i>"I didn't think there was any information about the negative charge... I didn't know what that meant, so I tried to go on the text."</i>

3.2.3.1 Theme One: Instruction Clarity

The theme of 'Instruction Clarity' was used to describe when participants either had difficulty interpreting the instructions or intentionally ignored the instructions. There was a tendency for many of the respondents to only scan the instructions without properly reading them. Furthermore, some respondents did not read the statement preceding a question in its entirety.

P1: "I picked A (Valid Assumption) first, then I re-read it and picked B (Invalid Assumption)"

I: "What made you change your mind?"

P1: "I am really bad at reading questions carefully and then I made sure I read it carefully."

Many respondents who exhibited P1's behaviour were quick to draw on outside knowledge. The Delphi report (Facione, 1990, p. 13) describes several affective dispositions of critical thinking when approaching questions or problems. These dispositions include 'diligence in seeking relevant information' and 'reasonableness in selecting and applying criteria'. Based on these dispositions alone, one might assume that this academic group were not disposed to critical thinking behaviour. However, the academics did not offer reasons as to why they did not read the instructions. The other element of this theme was the clarity of the instructions and providing relevant examples of what was meant by terms such as 'is of significant importance' or 'is not of significant importance'. Participants felt that any examples presented in the instructions provided little or no assistance in responding to the subsequent section. The level of engagement with the instructions highlighted that when administering the DOT-CCTT, the facilitator of the test needs to emphasize the importance of carefully reading the instructions, parent statements and questions. Furthermore, probing participants' engagement with the instructions of the test were identified as an important part of the content validity analysis, and specific lines of questioning were developed for the subsequent content validity study of the DOT-CCTTv2.

3.2.3.2 Theme Two: Wording of the Questions

The theme of 'Wording of the Questions' referred to evaluating the meaning of particular words within the questions or the parent statements. This theme encompassed how certain words were used in a scientific context. The wording of several questions led to confusion, causing the participants to draw on outside knowledge. For example, in question 22 clarification was suggested as individuals with scientific training might assume that they did not have enough information about the synthesis of an ester using a carboxylic acid or an anhydride described in the question:

P2: "I think it (Question 22) needs to say the same conditions. Same time."

Similarly, in question 25, clarification was required as scientific training led the participant to believe they didn't have all the information to be able to make a choice regarding the trails of natural product extraction:

P3: "It (Question 25) doesn't say the previous six trials- it doesn't say that it started with only those, there could have been many more before those six."

Unclear terminology at this stage prevented non-science participants (education developers) from attempting the questions, and was further compounded by the use of terms such as 'can only ever be', as exemplified in a discussion of diamagnetism and paramagnetism:

P4: "That's a big statement isn't it: 'only ever be'? Cause I don't know the chemistry, those words suggest there are no other possibilities."

In the same discussion, the use of the term 'rather than' elicited comments such as:

P3: "I got fooled cause it says 'paramagnetic rather than diamagnetic'"

P4: "it (Question 3) sounds like one or the other."

Academics were also confused, particularly chemists and science academics who defaulted to applying outside knowledge to these binary statements that they, in fact, knew to have more than two alternatives. As such, the tendency of the academics to focus on words

such as 'rather than' was looked at closely when writing the DOT-CCTTv2 and used sparingly only in questions that intended to convey a binary selection.

3.2.3.3 Theme Three: Information within the Statement

The theme of 'Information within the Statement' specifically referred to the participants' perceptions of the quality and depth of information provided in the parent statements used to respond to the questions. Several problems were identified within the statements preceding test questions. Comments suggested a disconnection between the statements and the corresponding test questions, for example in the question regarding the removal of zinc oxide in sunscreens:

P4: "From the first sentence-it was clearly non-sequitur. 'Zinc oxide protects against UV and should it be removed?' Well, why? It didn't say that. Unless other people know that zinc oxide is absorbed by the skin and that that's dangerous. It just says it does a really good job of reflecting back UV and so I couldn't answer any of those (questions) because I didn't believe in that first statement."

Participants such as P4 were not experts in chemistry and highlighted that this statement (used in Questions 8-10) drew on an assumed familiarity with the science:

P4: "It's a leap of faith which I don't know where science leads."

Similarly, there were discussions as to the relevance of bisphenol A in polycarbonate bottles (Questions 11-14). There was a prompt for analysing the arguments within the questions, but the participants felt the parent statement itself lacked context:

P3: "I would just flesh out the statement leading to the questions 11 to 14 with like you know 'there is some controversy about you know this chemical'"

P2: "There's no information about it getting to the DNA"

P1: "I don't know. I don't think I've got enough information"

P5: "Yeah exactly because (there's) not enough in the stem"

As P1 eluded, they did not have enough information to make a decision. The clarity of the instructions further compounded the problem for participants, as P5 suggested:

P5: "The instructions say its invalid if it doesn't relate statement or question, isn't supported by evidence and none of these (Questions 11-14) are in terms of what you've provided in the statement."

3.2.3.4 Theme Four: Prior Knowledge

The final theme 'Prior Knowledge' identified instances when participants had drawn on information not provided in the DOT-CCTTv1 to answer the questions. Several issues regarding prior knowledge emerged from the discourse. Some participants highlighted a sentiment of discomfort towards something outside their expertise, in this case P8 referring to the use of terms such as ester, carboxylic acid and anhydride in question 22:

P8: "The language is Egyptian (unfamiliar language) to me."

Other participants identified that there were some assumptions made about the use of the chemical notations. The extract below is in reference to questions 5 to 7 and the use of superscripts to denote positive and negative charges of ionic compounds.

P4: "I didn't think there was any information about the negative charge...I didn't know what that meant, so I tried to go on the text-so the only negative that I can see there is the two minus. There was nothing else about that-"

P6: "-And there is a one minus in the other one-that's what I used-but you have to have that-the knowledge of the symbolism being used."

Participants suggested that students may have prior general knowledge of a topic and expressed concerns that the questions were offering conflicting views. The example below referencing the removal of zinc oxide from sunscreen questions.

P7: *"This question (Questions 8-10), for example how sun screen works, what does it do. It's kind of general knowledge. They already know how a sun screen works. Maybe they're already using that knowledge."*

Finally some participants highlighted that having prior knowledge, specifically in science and/or chemistry, was to their detriment when attempting the questions. In reference to the discussion of the questions relating to ion-gated channels in a cell membrane (Questions 28-30) the following was highlighted:

P8: *"The idea that you can measure anything as five (mill molar), especially something like a-*

P9: *"But is that a problem with the question or your knowledge of accuracy?"*

When participants felt there was insufficient information in the statement or questions they were quick to draw on their outside knowledge. This often led to participants being very conservative in their responses to the questions. As highlighted above, participants thorough knowledge of precision and accuracy proved a significant hurdle to answering test questions.

P8: *(In response to Question22) "I didn't get the impression that it would fail completely. Just it's a probability thing and not about failure and success. You repeat this experiment a certain number of times and how many times does it, you know – throw these two molecules at each other a certain number of times and some of the time it succeeds or it fails it's like a scatter of cross section of physics."*

On occasion academics who were not chemistry experts, expressed anxiety when attempting questions they felt used assumed chemical knowledge:

P2: *"...I'm scared of trying...to offer an answer. Because we see it even in our biology students that the nearest mention of chemistry or maths they go 'argh, I can't answer that. I'm a biologist' – Umm. So it's that- it's that instinct to just say 'there is no possible way I could work this out'."*

In the above quote, P2 referred to their experience attempting questions regarding the diamagnetic and paramagnetic nature of organometallic complexes. P2 made a point that if an academic with their science experience was intimidated by the terminology used in the question, then it is fair to say many students would also have found the use of the terminology off-putting.

The discussion around prior knowledge prompted two key changes to the DOT-CCTT. The first was to clarify terms and notations which when writing the DOT-CCTTv1 were considered to be assumed knowledge. The most obvious example was the use of superscripts to denote charges of ionic species. While the DOT-CCTT was targeted at students with at least a high school understanding of science, it proved unwise to assume that all participants had the understanding of such chemical notation. For example, entry into the first year chemistry program at Monash University does not require a student to have completed chemistry studies at high school. For the DOT-CCTT to truly be independent of prior chemistry knowledge, specific notation such as charge superscripts were edited for the DOT-CCTTv2. Similarly, there was much discussion around the use of the terms 'diamagnetic' and 'paramagnetic'.

P5: "There are other types (of magnetism). Maybe this is one of the ones where having the knowledge makes the difference. Cause we know there are thing other than para(magnetic) and dia(magnetic) but if you don't know that-that 'rather than' suggests they are binary."

The principle of electron distribution within *d* subshells and the effect this has on magnetic properties of organometallic complexes is commonly taught in the first year of undergraduate chemistry courses. However, the theory is complex and nuanced which was problematic in the DOT-CCTTv1. P5 was very familiar with the field of organometallic complexes and they were very aware of the simplification of the terminology used throughout these questions. However, as previously mentioned, several academics from outside of the chemistry discipline found these questions intimidating with respect to the terminology, despite efforts to clarify the terminology in writing the parent statement and questions.

3.2.3.5 Content Validity Summary

A recurring discussion point within these focus groups was the concern regarding the limitations of measuring critical thinking using a multiple choice question tool. Academic participants commented in line with the literature (Bernard *et al.*, 2008; Butler, 2012). They highlighted that critical thinking is a complex cognitive process, and measuring the product of that process with binary right or wrong answers over-simplifies the critical thinking process, and potentially is an inaccurate measure of critical thought.

P8: “Can (you) actually test critical thinking via multiple choice? I’m not convinced because there are so many – I can come up with – you know as we have here, a million different ways to interpret questions...”

P8’s concern was a valid sentiment and, as described in Section 1.4, open-ended responses are more reflective of the cognitive processes involved in critical thinking (Halpern, 2016). However, large numbers of participants necessitated a multiple choice format which eliminates any marking variability (Ennis, 1993). Something that was implicit within the view regarding the limitation of multiple choice questions was the absence of any measure of critical thinking disposition. Displaying strong critical thinking skills on a test such as the DOT-CCTT does not suggest a person possesses strong critical thinking dispositions. The DOT-CCTT aimed to be able to determine expertise in critical thinking skills but not whether critical thinking behaviours would be elicited without explicit prompting. However as the DOT-CCTTv1 was completed on a voluntary basis, there appeared to be implications regarding measuring critical thinking disposition which emerged from the qualitative study of DOT-CCTTv2, as will be described in Section 3.3.4.

The themes generated from the analysis of the academic focus groups assisted in identifying problems with DOT-CCTTv1 questions which the majority of student participants answered incorrectly. However, due to time constraints the emphasis of the focus groups were to obtain qualitative feedback regarding the DOT-CCTTv1 and the academics did not undertake the DOT-CCTTv1 under test conditions, thus test scores were not collected. Furthermore, there was no opportunity to discuss the ‘Developing Hypotheses’ section with

the academics. It is important to note that the content validity was determined qualitatively and academics were not explicitly asked if they felt the overall test or specific questions measured critical thinking. Rather, the discussions which emerged in response to the questions on the DOT-CCTTv1 were used to identify issues the academics had in arriving at the correct responses. While four clear themes emerged from this study, a quantitative approach may have also proved useful. In the development of the Critical Thinking Assessment Task (CAT) (Stein *et al.*, 2007), researchers provided academics with a list of areas of critical thinking they believed the CAT assessed. The researchers then asked academics to identify whether the CAT in fact addressed these areas of critical thinking and which questions measured a component of critical thinking. A similar quantitative approach could be used in future validation studies of the DOT-CCTT however was not utilised in analysis of the DOT-CCTTv2 and DOT-CCTTv3.

3.3 DOT-CCTTv2

The second version of the DOT-CCTT (DOT-CCTTv2) addressed the feedback provided by the academic focus groups with particular attention paid to clarification of instructions, parent statements and very deliberate use of wording such as 'rather than'. In addition to these changes, chemical terminology was made explicit within the questions. For example super-scripts denoting ionic charges were explained.

The study of the DOT-CCTTv2 was interested in how reproducible student critical thinking scores were by measuring the test-retest reliability of the DOT-CCTTv2, and comparing performance on the DOT-CCTTv2 with performance on the WGCTA-S by measuring the convergent validity of the chemistry and generic tests. To continue to build on the content validity of the DOT-CCTT, student focus groups were also conducted whereby students discussed their responses to the WGCTA-S and the DOT-CCTTv2.

3.3.1 Demographic Data of the Participants

The participants for this cross-sectional study were comprised of equal numbers of male and female students, 15 participants identified English as their preferred language (75%),

two participants were bi-lingual, speaking English and French or Mandarin (10%), while three participants identified Chinese, Dari-Hazaragi or Vietnamese as their first language (15%). Their ages ranged from 18 to 21 with a median age of 19. Six students were undertaking first year chemistry courses (30%), five were taking second year courses (25%), seven were taking third year courses (35%), one taking fourth year (honours research) (5%), and one currently not studying any chemistry (5%). Four participants had completed third year chemistry courses (20%), six had completed second year courses (30%), seven had completed first year courses (35%) and three had completed studying chemistry in their final year of high school (15%).

Due to a small sample size of the cross-section of students, any statistical inferences were limited in their generalisability. Furthermore, it is important to acknowledge that the participants in this study were on average very high achieving and motivated students. Therefore, the study cannot be considered representative of the overall student populations the DOT-CCTT is targeting. 15 participants provided university entrance scores (ATAR) to be used as a measure of previous academic achievement. These scores ranged from the top 14.00 to 1.25 percentile with an average score in the 9.77 percentile. There is some discussion in the literature which suggests university entrance scores obtained in high school do not reflect intelligence and cognitive ability (Richardson, Abraham and Bond, 2012). However, a comparison of previous academic achievement, reported via ATAR scores, revealed a small positive correlation with scores obtained on the DOT-CCTTv2 ($\rho = .23$, $n = 15$, $p = .40$) and a moderately positive correlation with scores obtained on the WGCTA-S ($\rho = .47$, $n = 15$, $p = .08$).

3.3.2 Test-Retest Reliability

In total, 18 participants took part in test-retesting of the DOT-CCTTv2. A Wilcoxon signed rank test revealed no statistically significant change in score of the DOT-CCTTv2 due to test-retesting ($z = -.11$, $p = .91$) with a very small effect size ($r = .03$). The median score on the first day of testing (22.00) was similar to the median score on the second day of testing (22.50). The main concern regarding these findings was that the two attempts of the DOT-CCTTv2 were made on consecutive days. This was done in an attempt to reduce participant

attrition but risked participants responding exactly as they did in their first attempt of the DOT-CCTTv2 from memory. In fact, S9 identified that they felt they were remembering the answers from their previous attempt.

S9: *"The second time it felt like I was just remembering what I put down the day before."*

While the WGCTA-S manual did not provide any specific guidelines regarding the time interval between test attempts it did list test-retest studies conducted at different time intervals (Watson and Glaser, 2006, pp. 30-31). The manual listed three studies with test-retesting intervals of three months, two weeks or four days. Each of these studies reported test-retest correlations ranging from $p = .73$ to $.89$. The smaller p values were associated with shorter time frames between test-retesting. However, with respect to the DOT-CCTTv2, as the p value of the Wilcoxon's signed rank test was large enough ($.91$), it was unlikely that the DOT-CCTTv2 would have exhibited poor test-retest reliability were it to be administered over a longer time interval. To prevent the participants influencing each other's responses on the DOT-CCTTv2 the interviewers prevented discussion of the DOT-CCTTv2 during the first day of group interview. As suggested by S12, even the discussion of the WGCTA-S may have influenced how the group approached their second attempt of the DOT-CCTTv2:

S12: *"It would be interesting to see whether discussing the test that we did yesterday, the generalised one (WGCTA-S), whether you know, we spoke about some of the questions and how we stepped through with some of that logic. (I'd) Be curious to see whether that's - the way that we've understood other people analysing it has affected the way we've completed this test (DOT-CCTTv2) the second time....Because, you know... see(ing) how other people are stepping to through (questions) may influence you when you read these questions for the second time."*

While S12 raised an important point, test-retest reliability measurements suggested that this effect is not pronounced. Future test-retest reliability studies of the DOT-CCTT would best be better conducted over a three month interval, at the start and the end of a teaching semester.

Participant attrition becomes more of a concern due to the increased time interval but recruiting a larger number of participants, in excess of 100 students, would likely mitigate the issue of attrition.

3.3.3 Convergent Validity

Analysis of convergent validity was conducted using the WGCTA-S and the day one attempts of the DOT-CCTTv2, as there was no statistical significance between the scores of the two attempts of the DOT-CCTTv2. Additionally, day one attempts of the DOT-CCTTv2 were used to eliminate any possibility that the participants may have influenced each other's second attempt of the DOT-CCTTv2 as a result of discussing the WGCTA-S. The relationship between performance on the DOT-CCTTv2 and performance on the WGCTA-S was investigated using Spearman's ρ correlation to reveal a small positive correlation between the two variables, $\rho = .31$, $n = 18$, $p = .21$.

As stated in the WGCTA users guide (Watson and Glaser, 2006) the correlation with other reasoning tests should reflect the degree of similarity between the tests. In fact Pearson reports a range of correlations from .48 to .70 when comparing the WGCTA with other reasoning tests (Watson and Glaser, 2006, pp. 41-42). When a physics critical thinking test was validated against the Halpern Critical Thinking Assessment (HCTA) a correlation of $\rho = .45$ was determined at a statistically significant value ($p = .02$), despite the content of the two tests being different (Tiruneh *et al.*, 2016). The correlation of the physics critical thinking test and the HCTA was attributed to both tests focusing on similar critical thinking skills. As the WGCTA-S was presumed to be measuring critical thinking skills and the DOT-CCTTv2 was modelled after the WGCTA-S, it was initially assumed that the DOT-CCTTv2 measured the same aspects of critical thinking. However, the small correlation of the DOT-CCTTv2 relative to the WGCTA-S did suggest that the DOT-CCTTv2 was not necessarily measuring the same aspects of critical thinking as the WGCTA-S. While the correlation between the DOT-CCTTv2 and the WGCTA-S was not negative, it was not statistically significant ($p = .21$). The small positive correlation may have been due to the small number ($n = 18$) of self-selected

participants and did suggest the DOT-CCTTv2 exhibited some degree of convergent validity. However, as previously stated, a larger number of participants may provide more convincing data. For example, in the study of the physics critical thinking test, as few as 45 participants was sufficient to obtain statistically significant data (Tiruneh *et al.*, 2016). Convergent validity studies of future versions of the DOT-CCTT are recommended to be performed during teaching activities such as lectures, tutorials or laboratories in order to obtain a sufficient number of participants.

Several comparisons were made between the mean participant score of the WGCTA-S (mean = 30.4) and the norms provided in the accompanying WGCTA-S administrator's manual (Watson and Glaser, 2006). While norms were not specifically provided for undergraduate students, the WGCTA-S manual did provide norms by industry, occupation or position type/level. When the mean scores of the participants were compared by industry, the scores most closely aligned with education (Mean = 30.2). When considering the participants' mean score by occupation, the score most closely aligned with the accountant/auditor/bookkeeper category (Mean = 30.2). Finally the participants' scores most closely aligned with professional/individual contributor (Mean = 30.6) with respect to position type/level. As the WGCTA-S did not contain the norms specifically of higher education students or any definitions regarding the industries, occupations or position type/level, it was difficult to make any direct comparison. However, it was reassuring that the mean scores of the focus group were reflective of educated professionals and not at either extreme of the position type/level norms within the WGCTA-S manual. It would have been concerning had the average student in the focus group exhibited critical thinking reflective of business executive (Mean = 33.4) or of entry-level/hourly (Mean = 27.7) norms.

3.3.4 Content Validity

The qualitative analysis of the student focus group transcripts provided very useful data regarding the content validity of the DOT-CCTTv2. When discussing their responses within each focus group, the participants often arrived at a group consensus on the correct answers for both the DOT-CCTTv2 and the WGCTA-S. Rarely did the participants initially arrive at a

unanimous decision. In several instances on both tests, there were as many participants in favour of the incorrect response as there were participants in favour of the correct response. As will be discussed below, many of these stalemates resulted in discussions whereby participants re-evaluated their responses relative to their peers' point of view through peer learning.

The group interviews provided data which supported many of the decisions made regarding the design of the DOT-CCTTv2, such as changing the order in which sections appeared in the test and the details provided in the instructions. Overall the participants indicated they preferred the structure of the DOT-CCTTv2 over that of the WGCTA-S.

S9: *"The flow of the sections in the first test (DOT-CCTTv2) is clearer I think.*

The first (test) went from hypothesis and then...and then I think to conclusion."

S7: *"I found the questions a bit more interesting and engaging in general where as this one (WGCTA-S) seemed a bit more clinical. As was said previously it was more American so you go through some of the things you're like I don't know what I'm reading."*

S7 recognised the importance of the context of the tests and that they felt more comfortable doing the DOT-CCTTv2 as they attached greater significance to the chemistry content. Two participants did express their preference for the WGCTA-S, citing the detailed examples in the instructions of each section, and their frustration when attempting the DOT-CCTTv2, requiring them to recognise whether they were drawing on chemistry knowledge outside of the questions.

Participants also expressed their frustrations at having to respond to the questions on both tests using a multiple choice format. The participants felt restricted in their responses and that they would like to have clarified their reasoning either via written responses or in follow up interviews. Despite these issues, the participants recognised the pragmatic decision behind the use of multiple choice questions to evaluate large numbers of participants. However, some

participants did express that they would have been reluctant to engage with either the DOT-CCTTv2 or the WGCTA-S had long answer responses been required.

S13: "I don't want to have to write a small paragraph"

It would appear that the act of completing the DOT-CCTT on a voluntary basis may be reflective of the participants' 'care in focusing attention on the concern at hand' critical thinking disposition (Facione, 1990, p. 13). For example, the instructions were somewhat of a hurdle to overcome and nearly all participants stated how thoroughly they read the instructions on both tests. The participants also commented that their approach to the questions was influenced by the worked examples. Again, it was important to note that the students in this study were high achievers and it was not entirely surprising that they carefully read the instructions.

In total four themes emerged from the analysis of the transcripts which highlighted the approaches the participants took when attempting the tests and the challenges they encountered. These themes were 'Strategies for Completing the Tests', 'Difficulties Associated with Prior Knowledge', 'Awareness of Bias and Articulation of Critical Thought' and 'Evidence of Peer Learning'. A brief description and excerpts from the interviews can be found in Table 3. 3.

3.3.4.1 Theme One: Strategies for Completing the Tests

The theme of 'Strategies for Completing the Tests' describes both the participants' overall practice and increasing familiarity with the style of questions. The theme also described the specific cognitive techniques used in attempting to answer specific questions. The approach participants used when performing these tests was reflective of the fact that through repeated exposure to the tests, students became more familiar with the style of questions and their dependence on the examples provided diminished:

S1: "The examples were fairly detailed so you had an idea of what you should be doing when it came to the actual exercises."

S2: *"We've encountered them before. After doing a couple of them you're able to read through the questions and know what they wanted you to answer."*

The participants in the above statements described drawing on examples provided in the introductions of the tests and using repeated exposure to the style of questions to develop an understanding of how to demonstrate competency on those tests. For example, some

Table 3. 3 Themes identified in the qualitative analysis of the student focus groups for the DOT-CCTTv2 and the WGCTA-S

Theme	Description	Example
Strategies for Completing the Tests	Approaches participants took including dependence on examples, evaluating key words, construction of rules or hypothetical scenarios	<p><i>"...you could come back to it and then look at how each of the example questions were answered..."</i></p> <p><i>"Both the statement and the assumption are both ifs."</i></p> <p><i>"...it was really easy to think in terms of sets,"</i></p>
Difficulties Associate with Prior Knowledge	Participants consciously aware of their prior knowledge, either attempting to restrict its use, or it conflicted with their response to a given question	<p><i>"It's quite difficult to leave previous knowledge and experience off when you're trying to approach these (questions)."</i></p> <p><i>"...what do you define as a low yield?"</i></p>
Awareness of Bias and Articulation of Critical Thought	Evidence of critical thinking and use critical thinking terminology the participants were exposed to throughout the focus groups, in particular 'bias'	<p><i>"I think like the first section...was more difficult than the other because I think I had more bias in that question."</i></p> <p><i>"...after picking that up it falls outside of what we're looking at and becomes an unreasonable conclusion."</i></p>
Evidence of Peer Learning	Discourse between participants in which new insight was gain regarding how to approach test questions	<p><i>"To me, the fact that you know it starts talking about...fall outside of the information and is therefore an invalid assumption."</i></p> <p><i>"I don't think it's right now."</i></p>

participants had difficulty understanding what was meant by 'Assumption Made' and 'Assumption Not Made' in the 'Recognition of Assumption' section in the WGCTA-S and drew heavily on the worked examples provided in the introductions of the section. At the conclusion of this study, these participants had completed three critical thinking tests and were becoming familiar with how the questions were asked, and what was considered a correct response. However, test-retesting with the DOT-CCTTv2 indicated that there was no change in performance on the test.

The participants paying particularly close attention to the instructions in the DOT-CCTTv2 featured strongly within the theme of 'Strategies for Completing the Tests'. Participants read the instructions so carefully that they frequently used terminology found on the tests to express their thought processes. For example, students use of the term 'bias' featured heavily in their discussions. This observation highlighted the value of having robust instructions in the DOT-CCTT. However, there was concern that providing detailed instructions on the DOT-CCTT may in fact develop the participant's critical thinking skills in the process of attempting to measure it.

To investigate the influence detailed instructions and worked examples may have on performance on the DOT-CCTT, a study using a test group of students who attempt the DOT-CCTT without any instructions could be compared to the scores of a control group of students who receive instructions and examples. Heijltjes, van Gog, Leppink and Paas (2015) conducted such a study with 152 undergraduate economics students who were divided into six approximately equal groups ($n = 25-26$). All groups were required to complete a critical thinking skills test which focused on general knowledge syllogistic reasoning. Three groups were provided with explicit written instructions which provided step by step reasoning processes to apply to the test, while the other groups were provided with newspaper excerpts not directly related to the test. Participants who were exposed to the written instructions were found on average to perform 50% better on the critical thinking skills test as compared to those who did not receive written instructions at a statistical significance of $p < .001$. It does seem plausible that a similar phenomenon would occur with the DOT-CCTT, and evaluating the

impact of instructions and examples using control and test groups would be beneficial in future studies of the DOT-CCTT.

Other approaches included evaluation of words, categorising information and creating rules or hypothetical scenarios to interrogate the problems. The evaluation of words emerged as participants attributed meaning or significance to certain words.

S10: "This one was more about like-just reading off certain words. 'Cause like for (question on the WGCTA-S), if it was an 'or' instead of a 'nor' it would become a completely different question."

S3: "I always wonder whether I should pay serious attention to words like 'some' and 'always'. Like whether I should be really pedantic about it."

The Delphi report considers these behaviours part of the interpretation critical thinking skill which describes the ability 'to detect...relationships' or 'to paraphrase or make explicit....conventional or intended meaning' of a variety of stimuli (Facione, 1990, p. 8). Others consider this behaviour to be more reflective of problem solving skills, describing the behaviour as 'understanding of the information given' in order to build a mental representation of the problem (OECD, 2014, p. 31). When engaging with the WGCTA-S participants exhibited interpretation behaviours described in the Delphi report by evaluating the meaning of words such as 'some' and 'always'. Furthermore, when participants described their thinking with respect to the DOT-CCTTv2 they also explored the evaluation of more scientific terminology:

S4: "Well it (Parent statement on the DOT-CCTTv2) doesn't really mention the strength of attraction. It's only talking about what its formal charge is and how many things it can accept as a result."

In the above instance, the evaluation of the words 'attractions' and 'formal charge' featured heavily in S4's approach to the question. Conversely, there were unexpected discussions of the implication of the words 'lower yield' in a chemistry context and the relationship to a reaction failing. Participants pointed out underlying assumptions associated with terms such as yield in questions related to the synthesis of an ester using a carboxylic acid or an anhydride:

S4: *“Oh! What yields defined as. Cause if yield is just a mass of what you come up with over the mass of whatever you put in, then it won’t necessarily be a low yield.”*

S5: *“This higher probability of failure is an interesting one. What does it mean for a chemical reaction to fail?...does that mean that it’s going to – that the reaction will simply fail or does that simply mean that reaction will be – you know think of the kinetics – something like the reactions going to be slower or have a lower yield?”*

The comments made by S4 and S5 highlighted that the term yield was not necessarily reflective of obtaining a desired product and this produced frustration. Some participants highlighted that they viewed the term ‘failure’ from the perspective of a lay non-scientist without the understanding a chemist has of the term ‘yield’. As such, these observations were noted and the use of unintentionally ambiguous words were removed from the DOT-CCTTv3.

Another approach to the questions was to categorise the information to create rules or analogies which could then be tested.

S5: *“I’m thinking actual maths theorems where they have arrows going all different places and kind of thinking what things imply another. It’s pretty much exactly the same as reading this (WGCTA-S questions).”*

These behaviours combined what the Delphi experts refer to as part of interpretation and querying the evidence, the latter being part of inference (Facione, 1990, pp. 6-9). Once the participants felt they understood the information using their various representations, they proceeded to ‘formulate multiple alternatives for resolving a problem’ or in what can be described as representing and formulating (Facione, 1990, p. 9):

S9: *“You could say that there are lower taxes, there might not actually be enough funding to maintain the city.”*

S3: *"Maybe he got back problems and he went and saw chiro and that fixed his posture and hence fixed the general miasma of unhappiness and as a result of that he made new friends, not because of this (Dr) Baldwin guy."*

The second aspect of this theme was the application of problem solving skills and the generation of hypothetical scenarios whereby deductive logic could be applied:

S3: *"I find that with (Section) three, deduction, that it was really easy to think in terms of sets, it was easier to think in terms of sets rather than words, doodling Venn diagrams trying to solve these ones."*

What S3 described was an example of the participants explicitly categorising the information they were provided with in the parent statements and systematically analysing those relationships to answer the questions. These patterns of problem solving were most evident in the discussion of the WGCTA-S as compared to the discussions surround the DOT-CCTTv2, where participants had difficulty responding to the questions without drawing on the previous knowledge of chemistry and science. The possible alternatives drew on concepts outside of the information provided in the tests. It would appear that the use of knowledge outside the tests was difficult for participants to moderate. In fact the participants were more acutely aware of when they were at risk of using prior chemical knowledge as opposed to the use of other knowledge when attempting the WGCTA-S.

S10: *"It (question in the WGCTA-S) also never says what kind of tax. Because there's three kind of tax in America; income, property and sales. So if it has like low sales tax it probably has like really high property tax."*

S4: *"No I'm bias. That was my previous bias 'cause I studied BPA in the past so I didn't actually read the question properly"*

These comments illustrated the difficulties participants had with prior knowledge, which is a theme discussed below. These comments and those which appeared to demonstrate that prior knowledge seemed to play role in the generation of alternatives to problems, highlighted that

it would be useful to administer the DOT-CCTT to non-chemistry or non-science students and compare their performance against an equivalent group of chemistry students.

3.3.4.2 Theme Two: Difficulties Associated with Prior Knowledge

The second theme identified from this study, 'Difficulties Associated with prior Knowledge', described when participants drew on knowledge from outside the test in efforts to respond to the questions. In both the WGCTA-S and the DOT-CCTTv2, the instructions clearly stated to only use the information provided within the parent statements and the questions. These difficulties were most prevalent when participants described their experiences with the DOT-CCTTv2:

S4: "Being that it's (DOT-CCTTv2) coming from a chemistry background for all of us it's difficult to be able to leave behind some of that knowledge and be able to just answer those questions just from the information available."

For example, question 3 of the DOT-CCTTv2 asked participants to decide on the validity of an assumption regarding whether metal complexes can only ever be paramagnetic or diamagnetic based on the information provided. Below, S5's comments highlighted the conflict which arose from their prior chemistry studies when confronted with the restrictions of the DOT-CCTTv2:

S5: "Obviously it being paramagnetic is enough to deduce it has unpaired electrons but from the statement we do not have enough information to deduce that they're solely paramagnetic and diamagnetic"

Similarly, in question 5 the participants were asked to determine the validity of a statement regarding the relationship between the formal charge of anions and how readily anions accept hydrogen ions. In arriving at their answer, S3 drew on their outside knowledge of large molecules such as proteins to suggest:

S3: *"What if you had some ridiculous molecule that has like a 3 minus charge but the negative zones are all the way inside the molecule, then it would actually accept the H plus?"*

While S3's hypothesis led them to decide that the assumption was invalid, which was the correct response, the intended approach of this question was to recognise that the parent statement made no reference to how strongly cations and anions are attract to each other as exemplified by S6:

S6: *"I mainly chose invalid (assumption) because it didn't say whether or not protons are more attracted to CO₃ two minus"*

This theme was particularly prevalent in the 'Making Assumptions' section of the DOT-CCTTv2. This was due to the conflict which arose between the uncertainties allowed in making assumptions compared with the certainty of the scientifically verified facts the participants have learnt throughout their studies:

S5: *"An assumption it says (instructions for 'Making Assumptions' of the DOT-CCTTv2), an assumption can be implied..... so, one might assume that metals have something to do with the magnetism. That would probably be a valid assumption. It might be wrong, that's the thing about assumptions they're allowed to possibly be wrong."*

S7 provided a clear example of this conflict and acknowledged that their choice for a particular question was based on their previous knowledge:

S7: *"I got caught up in the fact that most of the time, yes that's the case scientifically but we don't have that information that we can actually make that assumption. But I did any way with my own personal bias."*

The theme of 'Difficulties Associated with Prior Knowledge' featured heavily throughout the groups. The theme was most prominent in students expressing their frustration and the mental effort required to identify what was chemical knowledge outside of the question, and therefore preventing that knowledge from influencing how they responded to the questions on

the DOT-CCTTv2. It was concerning that some participants felt they had to ‘un-train’ themselves of their chemical knowledge in order to properly engage with the DOT-CCTTv2. Some participants highlighted that they found the WGCTA-S easier as they did not have to reflect on whether they were using their prior knowledge. However, the issue around prior knowledge can be viewed in a positive light. Many participants described moments in which they had to reflect on their prior knowledge when attempting the DOT-CCTTv2:

S9: *“You had to think more oh am I using my own knowledge or what’s just in the question?”*

S9: *“I was like so what is assumed to be background knowledge. What’s background knowledge?”*

Comments such as those made by S9 demonstrated that participants were essentially asking themselves ‘why am I thinking what I’m thinking?’ which is indicative of high order meta-cognitive skills described by several critical thinking theoreticians (Facione, 1990, p. 10; Kuhn, 2000; Tsai, 2001). These students appeared to be questioning their responses to the DOT-CCTTv2 and whether their responses are based on their own pre-existing information or the information presented within the test.

Attempting and discussing the responses to the DOT-CCTTv2 also highlighted students’ misconceptions of chemistry concepts. Below, S9 described what they believed to be correct chemical theory being contradicted when asked about the relationship between the formal charge of an anion and its ability to accept hydrogen ions on the DOT-CCTTv2:

S9: *“For example this carbon – the carbonate question. Yeah of course there are gonna be more attractions but it didn’t really say about it, so I’m like how do I actually choose it.”*

S9 correctly identified that the DOT-CCTTv2 did not elude to a relationship, but S9 held on to the misconception that a greater negative formal charge always increased the attraction of hydrogen ions. No attempts were made by the interviewers during the focus groups to correct

the participants, however S9 correctly responded to the question on the DOT-CCTTv2 despite their chemical misconception. This observation highlights the teaching potential of a critical thinking test with a chemistry context. If the interviewer were to take the role of facilitator they could generate a group discussion underlying the critical thinking of the participants, who could hopefully begin to identify and correct their own misconceptions. For example, Garratt *et al.* (2000) provided students with chemistry critical thinking problems and students were required to work in small groups to come to a consensus regarding their reasoning. Alternatively, providing students with questions from the DOT-CCTT or questions derived from the DOT-CCTT in a tutorial style environment (Apple and Cutler, 1999; Jacob, 2004) may be beneficial to developing student critical thinking skills.

3.3.4.3 Theme Three: Awareness of Bias and Articulation of Critical Thought

The theme of 'Awareness of Bias and Articulation of Critical Thought' described the participants specifically applying the language from the instructions of the WGCTA-S and the DOT-CCTTv2 to articulate their thought processes in responding to the questions. In particular, participants were very aware of their prior knowledge referring to this as 'bias'. For example, below S8 was discussing how they had responded to question 17 in the 'Testing Hypothesis' section of the DOT-CCTTv2. In this particular question participants were required to decide if a statement regarding the expectations of trials to isolate a chemical were reasonable or unreasonable deductions. S8 identified that their choice was based on their prior knowledge rather than the information provided:

S8: *"I guess I put unreasonable, sorry reasonable deduction and I think it's because I may have brought some bias into this and I know that a lot of scientific discoveries are just serendipity."*

Similarly, in the discussion of the validity of arguments in questions 27 to 30 of 'Analysing Arguments' section of the DOT-CCTTv2, S4 acknowledges they did not follow the test instructions and drew on prior knowledge regarding bisphenol A (BPA):

S4: *"No I'm bias. That was my previous bias 'cause I studied BPA in the past so I didn't actually read the question properly"*

As a result of the participants recognising their prior science and chemistry knowledge as a source of potential bias, the participants sometimes attempted to view the questions from the perspective of individuals with limited scientific experience:

S4: *"I guess if you took out your bias and you were just reading it as a person who knew very little chemistry you would assume that failure means that it didn't form what you wanted it to form. So that's why I was like well that seems like a reasonable deduction to me"*

Here S4 discussed their response to question 16 in the 'Developing Hypothesis' section of the DOT-CCTTv2 with respect to probability of failure of an esterification reaction using a carboxylic acid compared to using an analogous anhydride. As a group there was much discussion of what was meant by the term 'failure' in the context of a chemical reaction. The group debated whether failure referred to the unsuccessful collisions at a molecular level or the absence of a product at the macroscopic level. As S4 suggested this level of detail could be considered a bias and responding to this question appropriately required restricting the use of detailed chemical knowledge.

These discussions also allowed the participants to refine the language they used when discussing their thought processes. For example S7 and S8 were discussing their approach to a question on the WGCTA-S which required information to be interpreted using some mathematical deduction:

S7: *"See its very mathematical logic type thing. And coming from a background of maths I reckon this is a bias that really helped."*

S8: *"It's not so much a bias but a skill."*

This is an example of several instances where the participants engaged in dialogues which helped refine the language they used in articulating their thoughts or helped them recognise

thinking errors. This was evident throughout the focus groups and describes the final emergent theme of 'Evidence of Peer Learning'.

3.3.4.4 Theme Four: Evidence of Peer Learning

The final theme of 'Evidence of Peer Learning' highlighted when participants discussed key terms or their thought processes which led to other participants recognising mistakes in their own thought processes and re-evaluating their approach to certain questions. For example, when discussing their thought processes regarding a question in the 'Deduction' section of the WGCTA-S, which required categorisation of groups, S3 and S9 engaged in the following discussion:

S9: *"What made it really hard is that they said you must assume that every statement is true and they said that all members enjoy playing and then that all members spend long hours practicing."*

S3: *"But the set of all symphonic members of the symphony orchestra does not include the set of all musicians."*

S9: *"Ah right, I didn't read that. I didn't even pick that up."*

Here S3 shared their strategy of having constructed sets or as they referred to their thought process as constructing Venn diagrams. Having constructed these sets they had correctly identified how elements of the question related whereas S9 had not. S9 recognising the connection they had initially failed to make and reconsidered their response in light of hearing S3's reasoning.

Similarly, when discussing question 20 in the 'Drawing Conclusions' section of the DOT-CCTTv2 participants re-evaluated their thinking. Question 20 focused on the correlation between the stability and reactivity of carbon-centred intermediates. S10 began to share their response to the question which generated some discussion in the group:

S10: *"I said unreasonable yesterday, I changed it to reasonable. If the intermediate is more stable it may be more reactive to certain types of reactions while less reactive to others."*

S2: *"It didn't really say anything about reactive or not-"*

S11: *"It says it the (re)activity -"*

S2: *"-is related."*

S11: *"-is related. It doesn't say it's like positively."*

S2: *"It doesn't actually say a direction."*

S10: *"Oh no! I messed up! I change mine back!"*

S2 and S11 discussed the nature of the correlation within the question and highlighted that the information provided in the parent statement regarding the correlation between reactivity and stability. They correctly identified that there was no indication as to whether the correlation was positive or negative. As this conversation was occurring, S10 realised they had made a mistake in their thought process, correcting themselves. Whilst not all discussions led to such overt recognition of errors in thinking, participants did state that engaging with their peers may have affected their own approach to the DOT-CCTTv2 on the second day of the study.

Evidence of the potential of the DOT-CCTT as a teaching tool was observed when the students displayed examples of peer learning (Cole and Wertsch, 1996; Schreiber and Valle, 2013). In these focus groups students put forward their thoughts regarding the questions, which were then analysed by their peers who provided their own perspectives and, in many instances, one or both of the students who engaged in the discourse, re-evaluated their responses. While most of these conversations were productive and generally resulted in a group consensus to arrive at the correct answers on both tests, there was concern regarding more vocal participants who insisted their views were correct and sometimes dominated the conversation. As previously mentioned the interviewers refrained from taking on the role of facilitator but if the DOT-CCTT were to be used as a teaching tool the risk of overly vocal

participants would need to be managed. The observation of peer learning behaviour further supports the possibility of incorporating DOT-CCTT style questions and discussions in teaching.

3.3.5 DOT-CCTTv2: Reliability and Validity Summary

Overall, the findings from the focus group study provided strong evidence that the DOT-CCTTv2 had good test-retest reliability and moderate convergent validity with respect to WGCTA-S. Limitations of the study were the small sample size of 20 participants and the short interval of 24 hours between attempts of the DOT-CCTTv2. The qualitative data collected from the focus groups suggested there is evidence DOT-CCTT style questions may be useful in peer learning environments. This finding warrants further investigation, perhaps by including chemistry critical thinking discussions into lecture, tutorial and laboratory setting then qualitatively analysing the discourses. If the DOT-CCTT questions were used as part of a teaching intervention, the students may become familiar with the questions and any quantitative analysis would require that a commercially available critical thinking test, such as the WGCTA-S be used to measure critical thinking unless additional DOT-CCTT style questions were developed for teaching interventions. Finally, the qualitative analysis of the focus groups identified opportunities to improve the content validity of the DOT-CCTT. These included clearer instructions with worked examples and further limiting the use of chemical terminology. As a result of these findings the DOT-CCTTv2 underwent significant changes to produce DOT-CCTTv3.

3.4 DOT-CCTTv3

The study of the DOT-CCTTv3 was particularly interested in the predictive criterion validity, which is a determination of a test being a valid predictor of future behaviour. To determine criterion validity, the DOT-CCTTv3 was administered to first year undergraduate, third year undergraduate, postgraduate chemistry students and academics with an interest in education. Additionally, there was sufficient data to be able to evaluate the internal reliability

using internal consistency (Cronbach's α), corrected item total correlations (CITC) for each section of the DOT-CCTTv3, and item difficulty analysis using the effect size of each question using Mann-Whitney U tests, similar to the study of DOT-CCTTv1.

3.4.1 Demographic Data of Participants

Table 3. 4 summarises the demographic data according to education group. The distribution of sex and age was representative of those observed for first year, third year and postgraduates. The distribution of sex and age for the Academics education group contained slightly more males (58%) and the median age (50) would suggest the majority of academics were mid to late career academics. The mean ATAR was reflective of the high admissions standards set by the two universities. Fewer postgraduates and academics provided an ATAR as many of these participants may not have completed their secondary education in Australia or before the ATAR was introduced (2009). 271 (91.6%) participants identified English as their preferred language with Chinese, Hebrew, Pashto, Portuguese, Punjabi, Russian, Spanglish, Swahili and Vietnamese being identified by one to two participants in each category.

Table 3. 4 Demographic data of participants who attempted the DOT-CCTTv3

Education Group	Mean ATAR	Female	Male	Median Age
First Year	87.10 ($n = 104$)	54% ($n = 64$)	43% ($n = 51$)	18 ($n = 117$)
Third Year	90.03 ($n = 55$)	39% ($n = 26$)	61% ($n = 41$)	20 ($n = 67$)
Postgraduates	87.35 ($n = 19$)	43% ($n = 19$)	57% ($n = 25$)	25 ($n = 44$)
Academics	89.58 ($n = 3$)	38% ($n = 15$)	58% ($n = 23$)	50 ($n = 37$)
Overall	88.23 ($n = 181$)	46% ($n = 124$)	52% ($n = 140$)	20 ($n = 283$)

3.4.2 Internal Reliability

The internal consistency data from Table 3. 5 shows the Cronbach's α determinations of the DOT-CCTTv3 as single scale or the DOT-CCTTv3 made up of five sub-scales were .71 and .78, respectively. These α values suggested the DOT-CCTTv3 viewed from either perspective exhibited acceptable internal reliability (DeVellis, 2012, p. 109), and could confidently be used to measure critical thinking skill as a single scale or the sum of five sub-scales. Upon calculating changes in α when questions/sub-scales were deleted it was found that from the single scale or five sub-scales perspective the DOT-CCTTv3 α values did not

improve if any questions or sub-scales were removed from the analysis (Appendix L Tables L4). The findings suggesting none of the questions or sub-scales were clearly measuring something other than one construct: critical thinking skills.

Table 3. 5 Summary of internal reliability data for the DOT-CCTTv3

	Sub-scales	Internal Consistency (Cronbach's α)	Mean CITC	Min CITC	Max CITC
DOT-CCTTv3 (Single scale)		0.71			
DOT-CCTTv3 (Five sub-scales)		0.78			
	Making Assumptions (Q1-7)	0.68	0.36	0.14	0.49
	Developing Hypotheses (Q8-13)	0.65	0.31	0.13	0.49
	Testing Hypotheses (Q14-18)	0.71	0.42	0.34	0.52
	Drawing Conclusion (Q19-23)	0.70	0.40	0.31	0.51
	Analysing Arguments (Q24-30)	0.71	0.42	0.29	0.53

The CITC and changes in α if questions were deleted for each sub-scale of the test further explored the internal consistency of the DOT-CCTTv3 (Appendix L Tables L6). The α values of each sub-scale presented in Table 3. 5 suggested that the questions from the 'Testing Hypotheses', 'Drawing Conclusions' and 'Analysing Arguments' sections possessed good internal consistency ($\alpha > .7$) and could confidently be added together to score their respective sub-scales. The 'Making Assumptions' and 'Developing Hypotheses' sections however exhibited low internal consistencies of .68 and .65 respectively. Table 3. 5 shows that on average the questions which make up each of the sub-scales of the DOT-CCTTv3 contribute to the measurement of their respective section as the mean CITC for each sub-scale was greater than .3. As can be seen from the minimum and maximum CITC of the 'Testing Hypotheses', 'Drawing Conclusions' and 'Analysing Arguments' sections show all questions in these sections exhibit a CITC $> .3$, and therefore contribute to the measurements of their respective sub-scales. The data from Table 3. 5 indicates that there are still questions which are not contributing to the score of their respective sub-scales as the CITC minimum

values for 'Making Assumptions' and 'Developing Hypotheses' are 0.14 and 0.13, respectively. When looking at the individual CITC, it appears questions 3 and 12 are the only questions which do not contribute to the overall score of their respective sub-scales ($CITC < .3$). Overall, the internal consistency and CITC data suggested DOT-CCTTv3 had acceptable internal reliability when viewed as single scale test made up 30 questions, or a test made up of five sections, each with an individual scale. The split-halves correlation, using the Spearman-Brown Coefficient, was found to have a value of $\rho_{cc'} = .80$, which added evidence that the DOT-CCTTv3 exhibited good internal reliability. The changes made to instructions, examples and wording of the DOT-CCTTv3 appear to have improved reliability as evidenced by the greater internal consistency of the DOT-CCTTv3 ($\alpha = .71$) relative to the DOT-CCTTv1 ($\alpha = .63$).

3.4.3 Item Difficulty Analysis

Figure 3. 4 shows the number of participants who answered each question on the DOT-CCTTv3 correctly, which was used to perform item difficulty analysis. With respect to the item difficulty of each question, the percentage of participants who answered the questions correctly was greater than the probability of answering the questions correctly by chance for every question on the DOT-CCTTv3.

Mann-Whitney U tests revealed all test questions to have a significant difference in the median score of participants who answered each DOT-CCTTv3 question correctly versus the median score of participants who answered the corresponding question incorrectly ($p < .001$). As can be seen in Figure 3. 5, questions 1, 14, 24 and 27 had large effect sizes ($r > .5$) (Cohen, 1988), suggesting that participants who answered these questions correctly were more likely to have obtained a higher score on the DOT-CCTTv3. Questions 2, 5, 9, 11, 15, 16, 17, 20, 21, 22, 25, 28, 29 and 30 had a medium effect size ($r > .3$) suggesting a moderate relationship between obtaining a high score on the DOT-CCTTv3 and answering these questions correctly. Questions 3, 4, 8, 13, 19, 23 and 26 had a small effect size ($r > .1$) indicating that answering

these questions correctly was not a strong indicator of obtaining high scores on the DOT-CCTTv3. Question 12 had a very small effect size ($r < .1$) suggesting that answering question 12 correctly was not an indicator of participants obtaining a high on the DOT-CCTTv3.

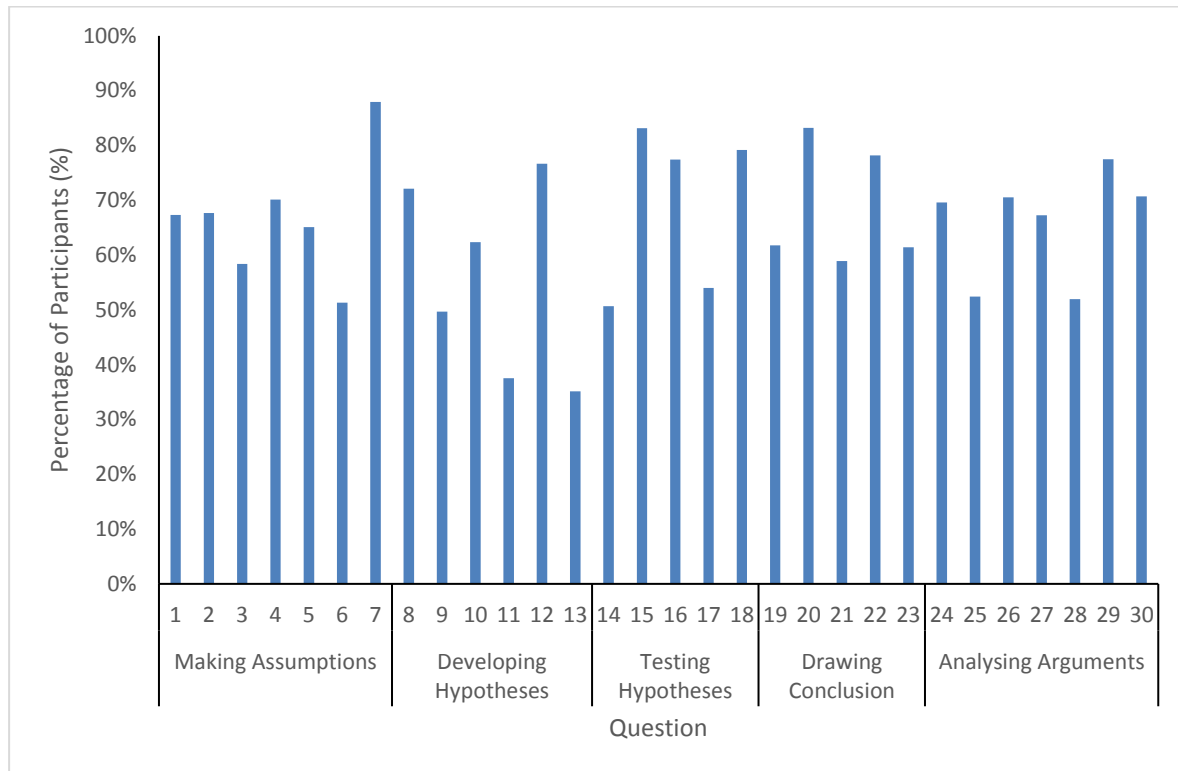


Figure 3. 4 Percentage of students who answered questions correct on DOT-CCTTv3

The data obtained from percentage of participants answering questions correctly and the effect size each question had on the overall test were useful indicators of the quality of each question. Mann-Whitney U tests revealed that 14 questions had a medium effect size ($r > .3$) and four questions had large effect size ($r > .5$). Comparing the effect sizes of the DOT-CCTTv3 questions to the DOT-CCTTv1, where only two questions had large effect sizes and nine question had medium effect sizes, demonstrated that there were a larger number of questions which were predictors of obtaining a high score on the DOT-CCTTv3.

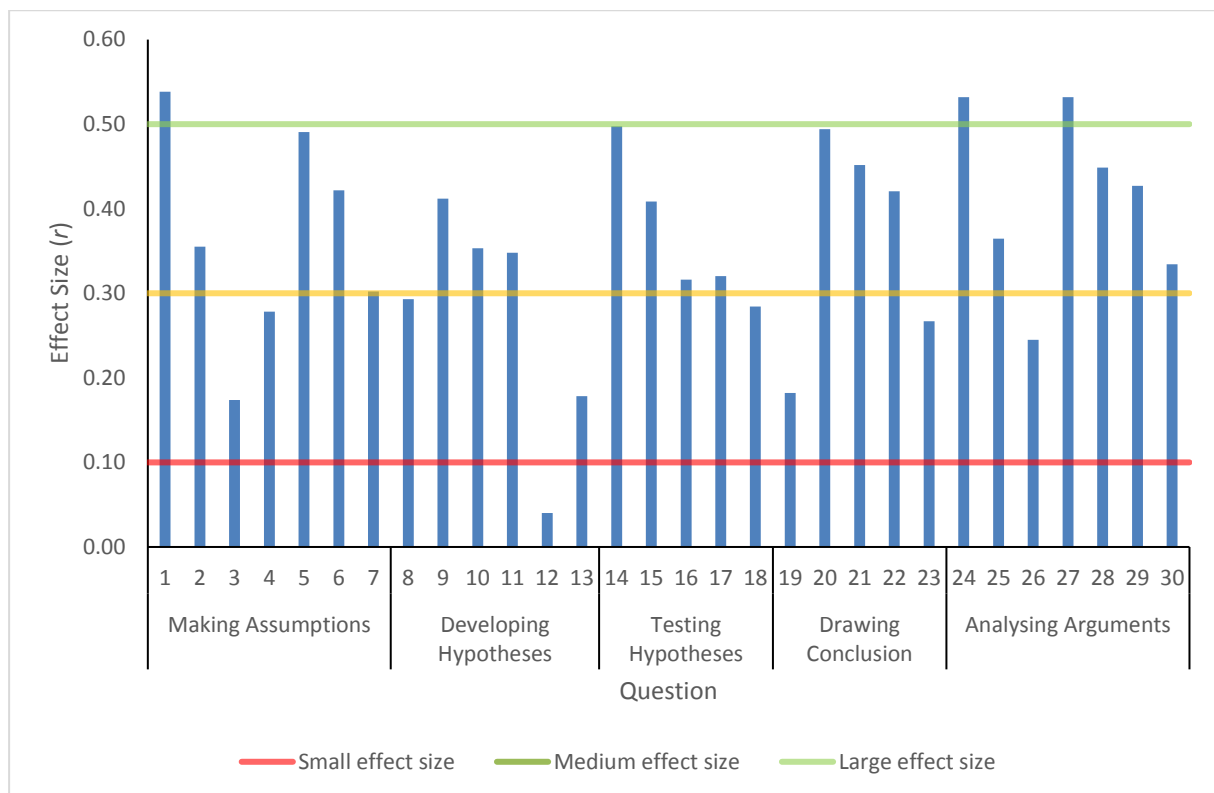


Figure 3. 5 Effect sizes (r) of the difference between median scores of participants who answered each DOT-CCTTV3 question correct or incorrect

With respect to assessing the difficulty of the questions, no questions were answered correctly by greater than 90% of participants, and all questions were answered correctly by participants more frequently than would be if the correct answer were obtained by chance. While neither the effect sizes nor the percentage of participants responding to questions correctly could be considered forms of reliability or validity, the data they provided was reassuring that all questions on the DOT-CCTTV3 were of an appropriate difficulty. Additional internal reliability could have been conducted using Rasch analysis which looks at a test question by question (Bond and Fox, 2007; Pallant and Tennant, 2007) and does not assume all test items are of equal difficulty. Instead, Rasch analysis determines the internal reliability of a test by plotting item difficulty against all participants' aptitude for that item and assesses how well this data correlates with the item response model (De Champlain, 2010). The strength of Rasch analysis is that it can identify questions which deviate from the item response model as they may be either too difficult or poorly constructed questions. Many proponents of Rasch analysis consider it to provide a truer representation of the internal reliability of a test (Bond

and Fox, 2007; De Champlain, 2010; Pallant and Tennant, 2007). Future versions of the DOT-CCTT would benefit from Rasch analysis as it would assist in identifying poor questions.

3.4.4 Criterion Validity

Several Mann-Whitney *U* tests were conducted to determine the criterion validity of the DOT-CCTTv3 (Table 3. 6). Significant differences in median scores were found between all education groups with the exception of the median scores of postgraduates compared to the median score of the academics. Of particular interest was that medium and large effect sizes were obtained when comparing the median scores of first years with respect to third years, postgraduates and academics. The findings from these analyses provided strong evidence that the DOT-CCTTv3 possessed good criterion validity when measuring the critical thinking skills of chemistry students up to and including post graduate students.

Table 3. 6 Mann-Whitney *U* tests comparing the median scores obtained on the DOT-CCTTv3 of each education group

	Education Group			
	1 st year (<i>n</i> = 119, <i>Md</i> = 16)	3 rd year (<i>n</i> = 67, <i>Md</i> = 21)	Postgraduates (<i>n</i> = 44, <i>Md</i> = 23.5)	Academic (<i>n</i> = 40, <i>Md</i> = 24)
1 st year		$p < .001, r = .39$	$p < .001, r = .64$	$p < .001, r = .59$
3 rd year	$p < .001, r = .39$		$p < .001, r = .30$	$p = .003, r = .28$
Postgraduates	$p < .001, r = .53$	$p < .001, r = .30$		$p = .691, r = .04$
Academic	$p < .001, r = .59$	$p = .003, r = .28$	$p = .691, r = .04$	

The comparison of the median scores on the DOT-CCTTv3 indicated the third year students, postgraduates and academics all performed statistically significantly ($p < .001$) better on the DOT-CCTTv3 relative to first year students. Similarly, postgraduates and academics performed statistically significantly ($p < .003$) better than third year students. The large effect sizes indicated that being a first year, third year, postgraduate or academic strongly effects likelihood of obtaining a high score on the DOT-CCTTv3. Interestingly, there appeared to be

no statistically significant difference in DOT-CCTTv3 scores when comparing postgraduates and academics ($p = .69$). If the assumption that critical thinking skill is correlated positively to years of tertiary study is valid, it is likely that the DOT-CCTTv3 was not sensitive enough to detect any difference in critical thinking skill between postgraduates and academics. Alternatively, the observation made earlier in this thesis that academics had a tendency to not read the instructions, parent statements and questions in full (Section 3.2.3.1) could have led to academics underperforming on the DOT-CCTTv3. There is also the possibility that the academic group and the postgraduates may, in fact, have the same level of critical thinking skill.

Table 3. 7 shows the percentage of questions answered correctly within each section of the DOT-CCTTv3 per education group. Table 3. 7 shows that every group answered the lowest percentage of questions correctly in the 'Developing Hypotheses' section. The first year students performed the best on the 'Drawing Conclusions' section , on average answering 60% of the questions correctly. The third year students answered 76% of questions correctly on both the 'Testing Hypotheses' and 'Drawing Conclusions' sections. The academics answered 83% of the 'Making Assumptions' section correctly, while the postgraduate group answered an average of 80% of questions correctly on all sections except 'Developing Hypotheses'. The academics completed the 'Testing Hypotheses' section almost as well as the 'Making Assumptions' section whilst on average answered the questions from the 'Drawing Conclusions' correctly 78% of the time and 'Assessment of Argument' correctly 74% of the time.

As can be seen in Table 3. 7, the percentage of questions answered correctly for each section of the test, and the percentage of total questions answered correctly increases with respect to years of tertiary study. Whilst the percentage of correct responses within each section varies slightly between the postgraduates and the academics, on average these groups answered the same percentage of questions correctly for the DOT-CCTTv3 overall. These trends for percentage of correct responses relative to years of study supported the hypothesis that participants who have undertaken more years of tertiary education would

obtain higher scores on the DOT-CCTTv3, thus adding evidence to suggest good criterion validity of the DOT-CCTTv3.

Table 3. 7 Percentage of questions answered correctly on the DOT-CCTTv3 according to education group

Education Group	Percentage of Question Answered Correctly of the DOT-CCTTv3					
	Section 1	Section 2	Section 3	Section 4	Section 5	DOT-CCTTv3
First Year (n = 119)	56%	47%	58%	60%	54%	55%
Third Year (n = 67)	74%	58%	76%	76%	73%	71%
Postgraduates (n = 44)	80%	65%	80%	80%	80%	77%
Academics (n = 40)	83%	68%	82%	78%	74%	77%
Overall (n = 270)	67%	55%	68%	68%	64%	64%

Descriptions of sections: 1 - Making Assumptions, 2 - Developing Hypotheses, 3 - Testing Hypotheses, 4 – Drawing Conclusions, 5 – Analysing Arguments.

Further evidence of criterion validity can be seen from Table 3. 8 to Table 3. 12, which show the percentage of participants who obtained the correct response for each question. From Table 3. 8 to Table 3. 12 there was a clear trend that the number of participants responding correctly to questions steadily increased as the level of education increased for questions 1, 4, 5, 6, 9, 12, 14, 15, 16, 19, 21, 27 and 30. Some questions appeared to plateau with respect to the percentage of participants selecting the correct response. The trends observed for these questions aligned well with the assumption that participants with more years of tertiary education possess greater critical thinking skill and suggest these questions exhibit good criterion validity.

Several question showed a steady improvement in DOT-CCTTv3 score relative to years of tertiary education then at some point exhibited a plateau effect. For example questions 3, 7, 10, 13, 15, 20 and 30 all appeared to plateau at the third year level, and questions 5, 9 and 28 appeared to plateau at a postgraduate level. These patterns suggested these questions exhibited good criterion validity up to and including the year level at which they plateau.

Finally, questions 2, 3, 5, 7, 11, 17, 18, 19, 22, 23, 24, 25, 26, 28 and 29 were answered correctly more frequently in the postgraduate group compared to the academics group. Of

these questions postgraduates obtained the correct responses for questions 11 and 17 at least 10% more often than academics. Also, question 19 was answered correctly less frequently by academics (58%) relative to the frequency of correct responses looking at the participants overall (62%).

Table 3. 8 Percentage of participants who answered questions 1 to 7 (Section 1: Making Assumptions) of the DOT-CCTTv3 correctly according to education group

	Question Number						
Education Group	1	2	3	4	5	6	7
First Year (n = 119)	52 %	51 %	54 %	65 %	50 %	39 %	81 %
Third Year (n = 67)	85 %	78 %	66 %	72 %	70 %	57 %	93 %
Postgraduates (n = 44)	77 %	84 %	66 %	77 %	90 %	68 %	95 %
Academics (n = 40)	90 %	78 %	65 %	80 %	88 %	75 %	93 %
Overall (n = 270)	67 %	68 %	58 %	70 %	65 %	51 %	88 %

Table 3. 9 Percentage of participants who answered questions 8 to 13 (Section 2: Developing Hypotheses) of the DOT-CCTTv3 correctly according to education group

	Question Number					
Education Group	8	9	10	11	12	13
First Year (n = 119)	63 %	35 %	52 %	31 %	75 %	31 %
Third Year (n = 67)	84 %	55 %	73 %	30 %	70 %	36 %
Postgraduates (n = 44)	73 %	72 %	73 %	66 %	73 %	36 %
Academics (n = 40)	83 %	75 %	73 %	45 %	93 %	40 %
Overall (n = 270)	72 %	50 %	63 %	38 %	77 %	35 %

Table 3. 10 Percentage of participants who answered questions 14 to 18 (Section 3: Testing Hypotheses) of the DOT-CCTTv3 correctly according to education group

	Question Number				
Education Group	14	15	16	17	18
First Year (n = 119)	38 %	69 %	68 %	45 %	73 %
Third Year (n = 67)	61 %	94 %	79 %	60 %	82 %
Postgraduates (n = 44)	64 %	95 %	82 %	73 %	91 %
Academics (n = 40)	70 %	98 %	95 %	63 %	80 %
Overall (n = 270)	51 %	83 %	77 %	54 %	79 %

Table 3. 11 Percentage of participants who answered questions 19 to 23 (Section 4: Drawing Conclusions) of the DOT-CCTTv3 correctly according to education group

	Question Number				
Education Group	19	20	21	22	23
First Year (n = 119)	63 %	76 %	42 %	66 %	57 %
Third Year (n = 67)	55 %	91 %	75 %	90 %	69 %
Postgraduates (n = 44)	64 %	95 %	79 %	95 %	70 %
Academics (n = 40)	58 %	95 %	83 %	88 %	63 %
Overall (n = 270)	62 %	83 %	59 %	78 %	61 %

Table 3. 12 Percentage of participants who answered questions 24 to 30 (Section 5: Analysing Arguments) of the DOT-CCTTv3 correctly according to education group

	Question Number						
Education Group	24	25	26	27	28	29	30
First Year (n = 119)	54 %	37 %	68 %	55 %	39 %	68 %	59 %
Third Year (n = 67)	80 %	44 %	81 %	75 %	60 %	92 %	79 %
Postgraduates (n = 44)	93 %	70 %	75 %	84 %	66 %	91 %	77 %
Academics (n = 40)	85 %	63 %	65 %	90 %	65 %	83 %	78 %
Overall (n = 270)	70 %	52 %	71 %	67 %	52 %	78 %	71 %

The academics obtaining the correct responses less frequently compared to other education groups was unexpected. Nevertheless, much like performance on the overall test, it was assumed that the percentage of academics answering a question correctly should be greater than the percentage of first years answering a given question correctly. While it is a fairly broad assumption that academics should outperform all groups on all questions, these results do highlight that there may be some issues regarding these particular questions. Question 11 will be discussed below as questions 11, 12 and 13 were all from the same parent statement, warranting their own discussion.

The lower percentage of correct responses from academics for questions 19 and 26 could potentially be attributed to engagement with the instructions and the use of prior knowledge in responding to these questions which could be investigated via interviews with academics. For example the qualitative study of DOT-CCTTv2 highlighted that students were very adept at identifying they had limited information regarding the correlation between a molecule's transition state and the molecule's stability. However, the application of typical thermodynamic principles taught in many general first year chemistry courses may lead to suggesting a positive correlation. Similarly in question 26, the argument of not removing zinc oxide from sunscreens was not directly related to decreasing risk of exposure to toxins. However, the instructions clearly stated that an argument is valid if it is important and related to the original question. The example in the instructions went so far as to use a potential cancer causing agent when describing a valid argument. Perhaps academics did not perceive the UV radiation-induced cancer as directly related to the argument and this may have led to the confusion. With respect to question 17, perhaps the higher proportion of incorrect responses could be attributed to engagement with the parent statement, and the academics not realising that there was no mention of the expectations regarding the amount of the chemical extracted.

Initially, the goal of the DOT-CCTT was to be able to discern between the critical thinking skills of undergraduate students for the purposes of assisting the development of critical thinking throughout their undergraduate degree. However, the implication that postgraduates and academics may exhibit the same level of critical thinking skills and the

possibility of a plateau effect with respect to the development of critical thinking skills could warrant further investigation. To determine whether, in fact, postgraduates and academics do exhibit the same degree of critical thinking skill, these groups could be asked to complete the WGCTA-S or a more complex written critical thinking test such as the Ennis-Weir Critical Thinking Essay Test (EWCTET) or the Halpern Critical Thinking Assessment (HCTA). If these tests were used and the median scores of academics were found to be statistically significantly higher than those of the postgraduates, then it would disprove the apparent plateau effect and suggest that the DOT-CCTTv3 is, in fact, not sensitive enough to discriminate between these groups. It also could be useful to conduct qualitative analyses of separate focus groups of postgraduates or academics, similar to those conducted in Section 2.12.4, investigating the questions where the percentage of postgraduate participants responding correctly to questions were equal to or greater than the percentage of academic participants responding to those same questions (See Table 3. 8 to Table 3. 12). Alternatively, qualitative data could be collected from interviews with postgraduates and academics using Live-scribe technology and think aloud protocols (Overton, Potter and Leng, 2013; Randles and Overton, 2015) while they attempt questions from the DOT-CCTTv3. The data that would be obtained from these approaches may highlight the similarities and/or differences the groups have in attempting these questions, looking at attention to instructions for example. Collecting data from academics had proven challenging throughout the development of the DOT-CCTT and future studies would best be conducted under conditions similar to those of other cohorts described within this thesis.

Questions 11, 12 and 13 warrant their own discussion as the number of participants who answered the questions correctly was rather low and/or the effect sizes of these questions were small. Questions 11 and 13 were answered correctly by less than 40% of participants. The effect sizes of question 12 and 13 were rather small ($r = .04$ and $.18$, respectively) in terms of the effect of answering these questions correctly had on predicting if the participant will obtain a high score on the DOT-CCTTv3. It is interesting that three questions from the same parent statement seem to be poor questions. In addition, these questions were drawn from the

'Developing Hypotheses' section which is the only section containing three multiple choice options.

With respect to question 11, the intention was that the participants identify that, although the passage stated that bird death sometimes occurs when they consumed plastic, the passage did not provide any information suggesting that some bird death will lead to some sea birds possibly becoming extinct, thus the inference was likely to be inaccurate. Academics and third year students predominately selected '*insufficient information to determine the accuracy*' possibly because they believed additional information was required and refrained from making decisions with incomplete data. Interestingly, the first year students predominately chose '*likely to be an accurate inference*' for question 11 possibly extrapolating the term 'sometimes even death' to mean extinction. However, a reluctance to act on incomplete information may not truly be the reason behind students' choice in question 11. First year students, third year students and postgraduates incorrectly decided that the correlation between the number of plastic coloured items and decline in the number sea birds (Question 13) was '*likely to be an inaccurate inference*', whilst academics correctly identified that there was insufficient information. While it initially appeared that question 13 was able to discriminate between the critical thinking skills of academics and that of other participants, the small effect size of this question suggested answering it correctly was not a strong predictor of a high score on the DOT-CCTTv3. Thus suggesting question 13 was not a reliable discriminator between academics and students. These observations may be due to the fact that the 'Developing Hypotheses' section contained three multiple choice options, that questions 11 to 13 perhaps need improvement in their wording, or that drawing inferences is an inherently more challenging critical thinking skill.

The explanation for the low effect size of question 12 was unclear. Over 75% of participants responded to the question correctly however the effect size was rather small ($r = .04$). The small effect size suggesting that answering question 12 correctly was not a predictor of obtaining a high score on the DOT-CCTTv3. Perhaps many participants had a strong pre-existing understanding regarding the relationship between consumer waste and

ocean pollution, and performance on the question was more dependent on prior or common knowledge rather than critical thinking skills. As suggested by an academic from the content validity study of DOT-CCTTv1:

P10: "They already know how a sun screen works. Maybe they're already using that knowledge."

Here P10 referred to the zinc oxide questions and participants potentially drawing on their prior knowledge with respect to the function of sun screen. The same may have been true of the questions relating to consumer waste and question 12 of the DOT-CCTTv3 would also benefit from think aloud protocol studies at undergraduate, postgraduate and academics level to better understand participants' reasoning behind their responses.

3.4.5 Discriminate Validity

There was a weak, positive correlation between score obtained on the DOT-CCTTv3 and previous academic achievement as measured by university entrance scores or ATAR scores ($\rho = .20$, $n = 194$, $p = .01$). This correlation suggested previous academic achievement was only a minor dependent with respect to score on the DOT-CCTTv3. This correlation was in line with previous observations collected during testing of the DOT-CCTTv2 where the correlation between previous academic achievement and performance on the test were found to be small ($\rho = .23$). Any conclusions regarding previous academic achievement and performance on the DOT-CCTTv2 were tentative due to the small sample size ($n = 15$). However, as the sample size used in the study of this relationship in the DOT-CCTTv3 was much larger ($n = 194$), the findings from the study of DOT-CCTTv3 suggested performance on the test was only slightly correlated to previous academic achievement.

A Mann-Whitney U test revealed no significant difference in the score ($U = 670.500$, $z = 1.835$, $p = .07$, $r = .22$) obtain on the DOT-CCTTv3 of 3rd year Monash University Chemistry students ($Md = 20$, $n = 44$) and 3rd year Curtin University Chemistry students ($Md = 22$, $n = 23$). As there was no statistically significant difference between scores of students from either university, the score obtained on the DOT-CCTTv3 was considered independent of where the

participant attended university. Whilst statistically speaking the results of the Mann-Whitney U analysis suggested the DOT-CCTTv3 performed well with students outside of Monash University, there was cause for concern as the p value of this analysis was approaching significance ($p < .05$). Were the Mann-Whitney U test to have revealed a statistically significant difference in scores between the universities, it would no longer be true that performance on the DOT-CCTTv3 was independent of which university the student attends. It is possible that an insufficient number of tests were completed, due to the opportunistic sampling from both universities, and obtaining equivalent sample sizes across several higher education institutions would better determine if the DOT-CCTTv3 performs outside of Monash University.

There was a large, positive correlation between score obtained on the DOT-CCTTv3 and age ($\rho = .50$, $n = 284$, $p < .01$). The Spearman's Rank-order correlation coefficient found a moderate positive correlation between age and score on the DOT-CCTTv3 for participants 30 years of age or less ($\rho = .43$, $n = 220$, $p < .01$), and a moderate positive correlation between age and score on the DOT-CCTTv3 for participants of 25 years of age or less ($\rho = .37$, $n = 199$, $p < .01$). It would appear that the score obtained on the DOT-CCTTv3 was dependent on the discriminant of age, and this dependence became stronger for test takers of 25 year or younger. The correlation at these restricted age limits suggested much development of critical thinking occurs at the undergraduate level. However, the effects of maturation may play a role as suggested by Macpherson and Owen (2010). It would be interesting to see if this pattern of development of critical thinking occurs at a high school level. Given the DOT-CCTT was designed to measure critical thinking independent of prior chemistry knowledge, it may be possible to administer the DOT-CCTT to final or penultimate year high school students. Alternatively the DOT-CCTT could be administered to a group of university students as a control group and a group of equivalent age and academic achievement who have not attended university. Pascarella (1999) administered the WGCTA-A to 30 college and 17 non-college participants using a test-retest approach and found that attending one year of university resulted in a 17% improvement in performance on the WGCTA-A. Performing a similar study

could assist in understanding if a similar trend in critical thinking skills occurs when learning chemistry and therefore clarify the relationship between age and score on the DOT-CCTT.

3.5 DOT-CCTTv3: Reliability and Validity Summary

Statistical analysis of the tests completed by first year and third year students, postgraduates and academics showed the DOT-CCTTv3 presented strong internal reliability, criterion validity, and discriminate validity with respect to previous academic achievement and where the student studies. Performance on the DOT-CCTTv3 appears to be correlated with age, which will require further exploration, possibly by administering the test in a cross-sectional study with non-university educated participants and determining any relationship between performance on the DOT-CCTTv3 and age. Issues arose from the study of the DOT-CCTTv3 included two questions which the majority of participants answered incorrectly and that there was no statistically significant difference between the DOT-CCTTv3 scores of postgraduates and academics. A qualitative analysis via think aloud protocols would be very useful in understanding how students, postgraduates and academics approach the DOT-CCTTv3, the similarities and differences between these groups, and how to alleviate these issues when developing the DOT-CCTTv4.

3.6 Implications for Practice and Further Work

The DOT-CCTTv3 showed evidence that it can discern between more and less experienced critical thinkers, suggesting the DOT-CCTTv3 is currently suitable for use in cross-sectional studies of participants drawn from first, second, third year or postgraduate chemistry student cohorts. Using the DOT-CCTTv3, it may be possible to evaluate the development of critical thinking across a degree program, like much of the literature which has utilised commercially available tests (Carter *et al.*, 2015). Using the DOT-CCTTv3 it may be possible to obtain base line data regarding the critical thinking skills of students and use this data to inform teaching practices aimed at developing critical thinking skills of students in subsequent years.

Whilst there is the potential to measure the development of critical thinking over a semester using the DOT-CCTTv3, there is evidence to suggest that a psychological construct, such as critical thinking, does not develop enough for measureable differences to occur in the space of a semester (Pascarella, 1999). While the DOT-CCTTv3 could be administered to the same cohort of students annually to form the basis of a longitudinal study, there are many hurdles to overcome in such a study regarding participant retention and the participants developing familiarity with the test. Maintaining a sufficiently large sample size could be remedied by starting with a large initial group of participants, such as all students from a course or subject, but concerns around participants developing familiarity with the DOT-CCTTv3 could be more difficult to resolve. Much like the CCTST and the WGCTA pre- and post-testing (Carter *et al.*, 2015; Jacobs, 1999), at least two versions of the DOT-CCTT may be required which still address the same critical thinking skills however the questions aren't necessarily exactly the same. Having a larger pool of questions does not prevent the participants from becoming familiar with the style of critical thinking questions which was demonstrated from the qualitative interviews with the Monash undergraduate chemistry students in Section 3.3.4.1. Furthermore, developing an additional test poses a host of issues and substantial research, such as validating the new questions and ensuring that all versions of the DOT-CCTT measure critical thinking with the same reliability and validity (Nunnally and Bernstein, 1994). As with commercially available critical thinking tests, the reliability and validity studies of the DOT-CCTT would be ongoing and newer, updated iterations of the test would need to be produced. To this end, cross-sectional studies are useful in identifying changes in critical thinking skills and the DOT-CCTTv3 has demonstrated it is sensitive enough to discern between the critical thinking skills of first year or third year undergraduate chemistry students.

The DOT-CCTT may be suited to actually develop critical thinking in teaching environments. The evidence from the focus group discussions of the DOT-CCTTv2 highlighted the potential for peer learning (See Section 3.3.4.4). The detailed instructions and the inclusion of examples appeared to provide students with a vocabulary with which to discuss their thought processes with their peers. Peer discussions are hardly recent innovations in attempting to

teach critical thinking (Apple and Cutler, 1999; Garratt *et al.*, 2000; Henderson, 2010). However, the goal and specific use of vocabulary has seldom occurred but does include research where educators have provided critical thinking questions to their students and explicitly taught syllogistic reasoning and informal logic within a chemistry course (Jacob, 2004).

The DOT-CCTT presents an opportunity for students to engage with critical thinking material that has been tailored to their interests and the rationale behind why these skills are relevant to the practice of science has been made explicit in the DOT-CCTT. Furthermore with adequate facilitation these questions could be used to identify chemical misconceptions and guide the students to recognise and correct their understanding collaboratively. While the DOT-CCTT as a whole may be too lengthy to discuss in its entirety in a single teaching session, were the DOT-CCTT to be broken up into its various sections, students could then respond to the questions individually followed by group discussions, much like the work conducted by Garratt *et al.* (2000). This approach implemented over several weeks could potentially enhance the students' understanding of critical thinking and the associated terminology.

The current administration of the DOT-CCTT as a voluntary test presents some interesting implications with respect to critical thinking disposition and representative sampling. Firstly, as the test is voluntary, students are self-selected, meaning that, on some level, they were possibly predisposed to thinking critically as they were willing to engage with the test. Studies reported in the literature typically use tests which were mandatory for participants to complete and offer little guidance with respect to disposition and voluntary test administration (Frisby and Traffanstedt, 2003). While the DOT-CCTT makes no claims to be able to evaluate the critical thinking disposition, the role of disposition must be taken into account when making generalisations regarding a students' critical thinking ability based on scores from the DOT-CCTT. Evidence for critical thinking disposition emerged in the qualitative discussion of the DOT-CCTTv2 where several participants indicated their reluctance to engage with the DOT-CCTTv2 were it to require essay style responses. In addition to these qualitative observations, an interesting phenomena was that many participants ceased attempting any

version of the DOT-CCTT when they were required to read slightly larger sections of text in the parent statements of the 'Developing Hypotheses' section. In fact, the DOT-CCTT may over estimate critical thinking skills and to alleviate this possibility, the option to complete the DOT-CCTT voluntarily could be removed, thus providing a more accurate measure of a students' critical thinking skills.

This thesis set out to ask whether a valid and reliable test could be designed to measure undergraduate chemistry students' critical thinking skills independent of extensive chemistry knowledge, while still being set within a broad chemistry context. To this end, the DOT-CCTT was constructed using literature definitions of critical thinking, the definitions of student, teaching staff and employers and commercially available tests. After three iterations of the DOT-CCTT, a wealth of quantitative and qualitative data demonstrated the reliability and validity of the test, with the expectation of weaker convergent validity, and performance being dependent on age. However, relative to commercially available tests, the DOT-CCTT is in its infancy and further reliability and validity studies are required.

The question still remains as to whether the DOT-CCTT is actually independent of an extensive chemistry knowledge, as evidenced throughout the interviews within this study. If at all, the data obtained throughout this research lends evidence to McPeak's (1981) notion that critical thinking cannot be independent of context. As evidenced through the qualitative discussions with students, and the performance of academics relative to first year students on the most recent version of the test, the best critical thinking does appear to occur within ones area of discipline expertise (Ennis, 1990).

To determine the dependency on chemical knowledge it would be useful to run cross-sectional studies much like the those conducted on the DOT-CCTTv3, however draw participants from outside the discipline of chemistry or even from outside the study of science. Alternatively, a cross-sectional study could be repeated in conjunction with an assessment using a chemistry concept inventory, which could then be used as a mediating variable, comparing chemistry concept scores and DOT-CCTT scores. Secondly, obtaining larger sample sizes of chemistry students outside of Monash University would help identify if the

DOT-CCTT does discern critical thinking skills of participants outside of Monash University. Thirdly, it would be useful to repeat the test-retest reliability, convergent validity and content validity focus groups using the DOT-CCTTv3 with a larger sample size and increase the duration of time between attempts of the DOT-CCTTv3. Finally, data collected from think aloud protocols in interviews where participants were asked to completed DOT-CCTT style questions could provide useful insight into the approaches used to answer these questions. In the near future higher education chemistry educators and researchers will be able to access the DOT-CCTTv3 and utilise it in their own teaching and research. These teachers and researchers will provide a wealth of data which will assist in reliability and validity studies of the DOT-CCTTv3 leading to improvement in the design of the DOT-CCTTv4.

Chapter 4 Conclusions

A chemistry critical thinking test (DOT-CCTT), which aimed to evaluate the critical thinking skills of undergraduate chemistry students at any year level irrespective of the students' prior chemistry knowledge, was developed over three versions (DOT-CCTTv1, DOT-CCTTv2 and DOT-CCTTv3). A qualitative analysis was initially conducted as part of a study published in 2017 in the peer-reviewed journal *Chemistry Education Research and Practice*, asking chemistry students, teaching staff and employers how they defined critical thinking. This analysis was performed as an exploratory study in an effort to understand the similarities and differences these groups had with respect to their conceptualisation of critical thinking and, therefore, to develop a test which best aligned with their definitions. The study found that all groups identified only deductive elements of critical thinking such as analysis and problem solving, with teaching staff and employers describing additional themes such as critique or context. All groups neglected to describe inductive logic elements such as judgement or inference. Consequently, it was decided that the DOT-CCTT would need to address a broader definition of critical thinking, more in line with the literature. The Watson Glaser Critical Thinking Appraisal (WGTCAT) was chosen as a model for the various iterations of the DOT-CCTT. Each version of the DOT-CCTT consisted of 30 multiple choice questions, divided into five sections: 'Making Assumptions', 'Developing Hypotheses', 'Testing Hypotheses', 'Drawing Conclusion' and 'Analysing Argument'. Each version of the test was evaluated for reliability and validity. Internal reliability, test-retest reliability, content validity, criterion validity, convergent validity and discriminate validity studies were conducted on various versions of the DOT-CCTT. The data obtained from these investigations were used to improve subsequent versions of the DOT-CCTT.

A total of 615 first year undergraduate chemistry students completed DOT-CCTTv1 in order to determine its internal reliability. Internal reliability was determined via internal consistency (Cronbach's α) to have a value of .34. For each of the 30 questions on the DOT-CCTTv1, 22 were answered correctly by 51-86% of participants, and eight of the 30 questions were answered correctly by less than 50% of participants, suggesting that none of

the questions were too easy and eight of questions were too challenging. Content validity of the DOT-CCTTv1 was determined via qualitative analysis of the recordings from group interviews of academics who had attempted the DOT-CCTTv1 to reveal four themes which highlighted changes required of the test. The themes of 'Instruction Clarity', 'Wording of the Question' and 'Information Within the Parent Statement' showed that additional information was required within the instructions and the parent statements of the test, and terms such as 'rather than' needed to be used with care. The theme of 'Prior Knowledge' highlighted that many of the chemical terms used in the DOT-CCTTv1 led to the academics either requiring chemical knowledge they did not possess or drawing on pre-existing knowledge from outside the test. The theme of 'Prior Knowledge' suggested these terms needed to be simplified or removed in future versions of the test. These four themes formed the basis of the changes made to the instructions, parent statements, wording and use of chemical terminology in writing the DOT-CCTTv2.

A cross-sectional group of 20 Monash University undergraduate chemistry students completed both the DOT-CCTTv2 and the WGCTA-S, and discussed both tests in recorded group interviews. The data obtained from the tests and recordings were used to determine test-retest reliability, convergent validity and content validity. The DOT-CCTTv2 showed good test-retest reliability, as determined by a Wilcoxon signed rank test, which revealed no statistically significant change in median scores as a result of attempting the DOT-CCTTv2 twice ($z = -.11$, $p = .91$, $r = .03$). Convergent validity was determined by comparing the median scores of the first attempt of the DOT-CCTTv2 and the median scores of the WGCTA-S using a Spearman's ρ correlation which revealed a positive correlation ($\rho = .31$, $n = 18$, $p = .21$) between test scores and suggested moderate convergent validity of the DOT-CCTTv2. Content validity was determined via qualitative analysis of the recording of the group interviews to reveal four themes regarding how the students approach the questions on the DOT-CCTTv2 and the WGCTA-S. The theme of 'Strategies for Completing the Tests' indicated the students were heavily reliant on the instructions and examples in both tests, and that they engaged in a variety of approaches such as evaluation of words, categorising information and creating

rules or hypothetical scenarios to interrogate the questions. The themes of 'Evidence of Peer Learning' and 'Awareness of Bias and Articulation of Critical Thought' revealed that discussion of DOT-CCTT style questions provided students with an opportunity to practice their critical thinking skills and suggests similar discussion activities may assist in developing students' with these skills. The theme of 'Difficulties Associated with Prior Knowledge' revealed students often had difficulty restricting the use of chemical knowledge not contained within the test, and assumptions associated with chemical terminology such as the term 'yield' led to students having difficulty in responding to questions on the DOT-CCTTv2. The data obtained from the themes of 'Strategies for Completing the Test' and 'Difficulties Associated with Prior Knowledge' led to changes in the DOT-CCTTv3 to include robust instructions, examples throughout the test, and the removal of apparently ambiguous terms such as 'yield'.

A cross-section of 296 first year undergraduate, third year undergraduate and postgraduate chemistry students, and academics from an online community of practice completed the DOT-CCTTv3. The analysis of these tests revealed the DOT-CCTTv3 possessed good internal reliability as measured by an internal consistency (Cronbach's α) value of .75. Discriminate validity was determined using Spearman's rank order correlation coefficient to find a large positive correlation between DOT-CCTTv3 score and age ($\rho = .50$, $n = 284$, $p < .01$), and a small positive correlation between DOT-CCTTv3 score and previous academic achievement as measure by ATAR score ($\rho = .20$, $n = 194$, $p = .01$). Additionally, a Mann-Whitney U test determined no significant difference in DOT-CCTTv3 scores of students who attended Monash or Curtin Universities ($p = .07$).

Criterion validity of the DOT-CCTTv3 was determined by comparing the median scores of first year undergraduate, third year undergraduate, postgraduate chemistry students, and academics from an online community of practice using Mann-Whitney U tests. The tests revealed that there was a statistically significant improvement in DOT-CCTTv3 scores relative to years of tertiary education, with the exception of the comparison between postgraduates and academics. The inability for the DOT-CCTTv3 to be able to discriminate between postgraduates and academics may be due to a lack of sensitivity of the DOT-CCTTv3,

insufficient engagement with the written instructions, or potentially that postgraduates and academics do indeed exhibit the same level of critical thinking skills. Overall the DOT-CCTTv3 appears to demonstrate reliability and validity when measuring the critical thinking skills of an undergraduate chemistry student cohort.

Further investigation of the DOT-CCTTv3 with respect to test-retest reliability, convergent validity, discriminate validity and content validity would be useful in further improving the reliability and validity of the test. Test-rest reliability and convergent validity studies will be conducted by administering the DOT-CCTTv3 and the WGCTA-S to all students from a subject or course, such as first year chemistry at the start of a teaching semester, with the DOT-CCTTv3 administered again at the end of the teaching semester, using the same cohort of students. With a larger number of participants and an interval of approximately three months between DOT-CCTTv3 attempts, it is anticipated that no significant change in scores on the DOT-CCTTv3 will be observed, and a stronger correlation between scores on the DOT-CCTTv3 and WGCTA-S will be obtained, relative to the convergent validity study conducted on the DOT-CCTTv2.

Discriminates of further interest include age and the institution the students attend, and their area of study. With respect to the discriminate of age, there is the potential to administer the DOT-CCTTv3 with a cross-section of participants who have not attended university and use this data in conjunction with the existing data of the DOT-CCTTv3 to determine if performance on the DOT-CCTTv3 is dependent on age. In terms of where the student attends university and what area they study, the DOT-CCTTv3 will be administered to Monash University students outside of the School of Chemistry, such as the Schools of Physics and Astronomy or Biological Sciences; to students outside of the Faculty of Science, for example within the Faculties of Medicine, Nursing and Health Sciences or Business and Economics; and students at other higher education institutions around the world. If the DOT-CCTTv3 is truly a measure of critical thinking independent of prior chemistry knowledge, scores obtained on the DOT-CCTTv3 should have no dependence on what the student studies and should be comparable to the test scores of students from equivalent years of undergraduate study.

Similarly, DOT-CCTTv3 scores obtained from chemistry students around the world should be similar to those of chemistry students at Monash University, adding supporting evidence that the DOT-CCTTv3 measures critical thinking skills independent of where the student cohort has studied. Finally content validity data will be collected and qualitatively analysed via focus group discussions of undergraduates, postgraduates and academics who have attempted the DOT-CCTTv3, similar to those described throughout this thesis. A detailed understanding of how of undergraduates, postgraduates and academics approach questions on the DOT-CCTTv3 and the WGCTA-S will be collect from individual interviews using think aloud protocols while participants attempt questions from both tests. The data obtained from all of the listed potential studies will add reliability and validity data to the DOT-CCTTv3 and provide valuable insight into development of the DOT-CCTTv4.

Assisting university chemistry educators and researchers in the development of their students' critical thinking skills was the original purpose of designing a valid and reliable test to measure undergraduate chemistry students' critical thinking skills independent of extensive chemistry knowledge, while still set within a broad chemistry context. To this end the DOT-CCTTv3 offers a tool with which to measure a student's critical thinking skills and the effect of any teaching interventions specifically targeting the development of critical thinking. The test is suitable for studying the development of critical thinking using a cross-section of students, and may be useful in longitudinal studies of a single cohort, though there are concerns of participants developing familiarity with the test. Qualitative data from this study suggests that use of the test itself may be beneficial in the development of student critical thinking. Discussion groups set in teaching environments where students discuss critical thinking questions similar to those of the DOT-CCTT, may provide students with the opportunity to practice their critical thinking skills.

In summary, research reported within this thesis provides a body of evidence regarding reliability and validity of the DOT-CCTT, and that the DOT-CCTT offers the chemistry education community a valuable research and education tool with respect to the development of undergraduate chemistry students' critical thinking skills.

Chapter 5 References

- Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, **85**(2), 275-314.
- Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta- analysis. *Review of Educational Research*, **78**(4), 1102-1134.
- ACT. (2017). *Act CAAP technical handbook*. Iowa, USA: ACT. Retrieved from <http://www.act.org/content/dam/act/unsecured/documents/CAAP-TechnicalHandbook.pdf>, Accessed on 07/09/2017.
- Apple, T., & Cutler, A. (1999). The rensselaer studio general chemistry course. *Journal of Chemical Education*, **76**(4), 462-463.
- AssessmentDay Ltd. (2015). *Watson Glaser critical thinking appraisal*. Retrieved from <https://www.assessmentday.co.uk/watson-glaser-critical-thinking.htm>, Accessed on 03/07/2015.
- Australian National University. (2015). *Chemistry major*. Retrieved from <http://programsandcourses.anu.edu.au/2016/major/CHEM-MAJ>, Accessed on
- Badcock, P. B. T., Pattison, P. E., & Harris, K.-L. (2010). Developing generic skills through university study: A study of arts, science and engineering in Australia. *Higher Education: The International Journal of Higher Education and Educational Planning*, **60**(4), 441-458.
- Bailin, S. (2002). Critical thinking and science education. *Science & Education*, **11**, 361–375.
- Barnett, R. (1997). *Higher education: A critical business*. Buckingham: Open University Press.
- Behar-Horenstein, L. S., & Niu, L. (2011). Teaching critical thinking skills in higher education: A review of the literature. *Journal of College Teaching & Learning*, **8**(2), 25-41.

- Bennett, S. W. (2008). Problem solving: Can anybody do it? *Chemistry Education Research and Practice*, **9**(1), 60-64.
- Bernard, R. M., Zhang, D., Abrami, P. C., Sicol, F., Borokhovski, E., & Surkes, M. A. (2008). Exploring the structure of the Watson-Glaser critical thinking appraisal: One scale or many subscales? *Thinking Skills and Creativity*, **3**(1), 15-22.
- Biggs, J. (2012). What the student does: Teaching for enhanced learning. *Higher Education Research & Development*, **31**(1), 39-55.
- Blattner, N. H., & Frazier, C. L. (2002). Developing a performance-based assessment of students' critical thinking skills. *Assessing Writing*, **8**(1), 47-64.
- Bond, T. G., & Fox, C. M. (2007). *Applying the rasch model : Fundamental measurement in the human sciences*. Mahwah, NJ: Mahwah, NJ : Lawrence Erlbaum Associates.
- Bryman, A. (2008). *Social research methods* (3rd ed.). Oxford: Oxford University Press.
- Butler, H. A. (2012). Halpern critical thinking assessment predicts real-world outcomes of critical thinking. *Applied Cognitive Psychology*, **26**(5), 721-729.
- Carmel, J. H., & Yezierski, E. J. (2013). Are we keeping the promise? Investigation of students' critical thinking growth. *Journal of College Science Teaching*, **42**(5), 71-81.
- Carter, A. G., Creedy, D. K., & Sidebotham, M. (2015). Evaluation of tools used to measure critical thinking development in nursing and midwifery undergraduate students: A systematic review. *Nurse Education Today*, **35**(7), 864-874.
- Chapman, E., & O'Neill, M. (2010). Defining and assessing generic competencies in Australian universities: Ongoing challenges. *Education Research and Perspectives*, **37**(1), 105-125.
- Chase, A. M., Clancy, H. A., Lachance, R. P., Mathison, B. M., Chiu, M. M., & Weaver, G. C. (2017). Improving critical thinking via authenticity: The caspie research experience in a military academy chemistry course. *Chemistry Education Research and Practice*, **18**(1), 55-63.
- Cloonan, C. A., & Hutchinson, J. S. (2011). A chemistry concept reasoning test. *Chemistry Education Research and Practice*, **12**, 205-209.

- Cohen, J. W. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cole, M., & Wertsch, J. V. (1996). Beyond the individual-social antinomy in discussions of piaget and vygotsky. *Human Development*, **39**(5), 250-256.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, **16**, 297-334.
- Daly, W. M. (2001). The development of an alternative method in the assessment of critical thinking as an outcome of nursing education. *Journal of Advanced Nursing*, **36**(1), 120-130.
- Danczak, S. M., Thompson, C. D., & Overton, T. L. (2017). What does the term critical thinking mean to you? A qualitative analysis of chemistry undergraduate, teaching staff and employers' views of critical thinking. *Chemistry Education Research and Practice*, **18**(3), 420-434.
- Davies, M. (2006). An 'infusion' approach to critical thinking: Moore on the critical thinking debate. *Higher Education Research & Development*, **25**(2), 179-193.
- Davies, M. (2013). Critical thinking and the disciplines reconsidered. *Higher Education Research & Development*, **32**(4), 529-544.
- De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, **44**(1), 109-117.
- Desai, M. S., Berger, B. D., & Higgs, R. (2016). Critical thinking skills for business school graduates as demanded by employers: A strategic perspective and recommendations. *Academy of Educational Leadership Journal*, **20**(1), 10-31.
- DeVellis, R. F. (2012). *Scale development : Theory and applications* (3rd ed. ed.). Thousand Oaks, Calif.: Thousand Oaks, Calif. : SAGE.
- Domin, D. (1999). A content analysis of general chemistry laboratory manuals for evidence of higher-order cognitive tasks. *Journal of Chemical Education*, **76**(1), 109-112.
- Drennan, J. (2010). Critical thinking as an outcome of a master's degree in nursing programme. *Journal of Advanced Nursing*, **66**(2), 422-431.

- Dressel, P. L., & Mayhew, L. B. (1954). *General education: Explorations in evaluation*. Washington, D.C.: American Council on Education.
- Edwards, D., Perkins, K., Pearce, J., & Hong, J. (2015). *Work intergrated learning in STEM in Australian universities*. Retrieved from http://www.chiefscientist.gov.au/wp-content/uploads/ACER_WIL-in-STEM-in-Australian-Universities_June-2015.pdf, Accessed on 05/12/2016.
- Ennis, R. H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational Researcher*, **18**(3), 4-10.
- Ennis, R. H. (1990). The extent to which critical thinking is subject-specific: Further clarification. *Educational Researcher*, **19**(4), 13-16.
- Ennis, R. H. (1993). Critical thinking assessment. *Theory Into Practice*, **32**(3), 179-186.
- Ennis, R. H., & Chatten, G. S. (2015). *An annotated list of english-language critical thinking tests*. Retrieved from www.criticalthinking.net/TestListDraft050315.docx, Accessed on 19/07/2017.
- Ennis, R. H., & Weir, E. (1985a). *The Ennis-Weir critical thinking essay test*. Retrieved from http://faculty.education.illinois.edu/rhennis/tewctet/Ennis-Weir_Merged.pdf, Accessed on 03/09/2015.
- Ennis, R. H., & Weir, E. (1985b). *The Ennis-Weir critical thinking essay test: Test, manual, criteria, scoring sheet*. Retrieved from http://faculty.education.illinois.edu/rhennis/tewctet/Ennis-Weir_Merged.pdf Accessed on 09/10/2017.
- Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Executive summary. "The delphi report"*. Millbrae, CA: T. C. A. Press
- Facione, P. A., Sánchez, C. A., Facione, N. C., & Gainen, J. (1995). The disposition toward critical thinking. *The Journal of General Education*, **44**(1), 1-25.

- Ferguson, R. L. (2007). Constructivism and social constructivism. In G. M. Bodner & M. Orgill (Eds.), *Theoretical frameworks for research in chemistry and science education*. Upper Saddle River, NJ, United States: Pearson Education (US).
- Flynn, A. B. (2011). Developing problem-solving skills through retrosynthetic analysis and clickers in organic chemistry. *Journal of Chemical Education*, **88**, 1496-1500.
- Frisby, C. L., & Traffanstedt, B. K. (2003). Time and performance on the california critical thinking skills test. *Journal of College Reading and Learning*, **34**(1), 26-43.
- Frye, B., Alfred, N., & Campbell, M. (1999). Use of the Watson-Glaser critical thinking appraisal with bsn students. *Nursing and health care perspectives*, **20**(5), 253-255.
- Gadzella, B., Hogan, L., Masten, W., & Stacks, J. (2006). Reliability and validity of the Watson-Glasere critical thinking appraisal-forms for different academic groups*. *Journal of Instructional Psychology*, **33**(2), 141-143.
- Garratt, J., Overton, T., & Threlfall, T. (1999). *A question of chemistry*. Essex, England: Pearson Education Limited.
- Garratt, J., Overton, T., Tomlinson, J., & Clow, D. (2000). Critical thinking exercises for chemists. *Active Learning in Higher Education*, **1**(2), 152-167.
- Giancarlo, C. A., & Facione, P. A. (2001). A look across four years at the disposition toward critical thinking among undergraduate students. *The Journal of General Education*, **50**(1), 29-55.
- Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist*, **39**(2), 93-104.
- Graduate Careers Australia. (2015). *Chemistry - bachelor graduates (all)*. Retrieved from <http://www.graduatecareers.com.au/Research/GradJobsDollars/BachelorAll/Chemistry/index.htm>, Accessed on 15/04/2015.
- Gupta, T., Burke, K. A., Mehta, A., & Greenbowe, T. J. (2015). Impact of guided-inquiry-based instruction with a writing and reflection emphasis on chemistry students' critical thinking abilities. *Journal of Chemical Education*, **92**(1), 32-38.

- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, **8**(1), 23-34.
- Halpern, D. F. (1993). Assessing the effectiveness of critical thinking instruction. *The Journal of General Education*, **50**(4), 238-254.
- Halpern, D. F. (1996a). Analyzing arguments. In D. F. Halpern (Ed.), *Thought and knowledge : An introduction to critical thinking* (3rd ed., pp. 167-211). Mahwah, N.J.: L. Erlbaum Associates.
- Halpern, D. F. (1996b). *Thought and knowledge : An introduction to critical thinking* (3rd ed.). Mahwah, N.J.: L. Erlbaum Associates.
- Halpern, D. F. (1998). Teaching critical thinking for transfer across domains. Dispositions, skills, structure training, and metacognitive monitoring. *The American psychologist*, **53**, 449-455.
- Halpern, D. F. (2016). *Manual: Halpern critical thinking assessment* Retrieved from https://drive.google.com/file/d/0BzUoP_pmwy1gdEpCR05PeW9qUzA/view Accessed on 09/10/2017.
- Hambur, S., Rowe, K., & Luc, L. T. (2002). *Graduate skills assessment stage one validity study*. Retrieved from https://research.acer.edu.au/higher_education/27/, Accessed on 19/01/2015.
- Hassan, K. E., & Madhum, G. (2007). Validating the Watson Glaser critical thinking appraisal. *Higher Education*, **54**(3), 361-383.
- Heijltjes, A., van Gog, T., Leppink, J., & Paas, F. (2015). Unraveling the effects of critical thinking instructions, practice, and self-explanation on students' reasoning performance. *Instructional Science*, **43**(4), 487-506.
- Henderson, D. E. (2010). A chemical instrumentation game for teaching critical thinking and information literacy in instrumental analysis courses. *Journal of Chemical Education*, **87**, 412-415.
- Huber, C. R., & Kuncel, N. R. (2016). Does college teach critical thinking? A meta-analysis. *Review of Educational Research*, **86**(2), 431-468.

- Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence : An essay on the construction of formal operational structures*. London: Routledge & Kegan Paul.
- Insight Assessment. (2013). *California critical thinking skills test (CCTST)*. Request information. Retrieved from <http://www.insightassessment.com/Products/Products-Summary/Critical-Thinking-Skills-Tests/California-Critical-Thinking-Skills-Test-CCTST>, Accessed on 07/09/2017.
- Insight Assessment. (2017). *California critical thinking disposition inventory (CCTDI)*. Retrieved from <https://www.insightassessment.com/Products/Products-Summary/Critical-Thinking-Attributes-Tests/California-Critical-Thinking-Disposition-Inventory-CCTDI>, Accessed on 07/09/2017.
- Iwaoka, W. T., Li, Y., & Rhee, W. Y. (2010). Measuring gains in critical thinking in food science and human nutrition courses: The Cornell critical thinking test, problem-based learning activities, and student journal entries. *Journal of Food Science Education*, **9**(3), 68-75.
- Jackson, D. (2010). An international profile of industry-relevant competencies and skill gaps in modern graduates. *international journal of management education*, **8**(3), 29-58.
- Jacob, C. (2004). Critical thinking in the chemistry classroom and beyond. *Journal of Chemical Education*, **81**(8), 1216-1223.
- Jacobs, S. S. (1999). The equivalence of forms A and B of the California critical thinking skills test. *Measurement and Evaluation in Counseling and Development*, **31**(4), 211-222.
- Johnson, R. H., Blair, J. A., & Hoaglund, J. (1996). *The rise of informal logic : Essays on argumentation, critical thinking, reasoning, and politics*. Newport, Va.: Newport, Va. : Vale Press.
- Jones, S., Yates, B., & Kelder, J.-A. (2011). *Learning and teaching academic standards project, science: Learning and teaching academic standards statement september 2011*. Retrieved from <http://www.acds-tlcc.edu.au/wp->

content/uploads/sites/14/2015/02/altc_standards_SCIENCE_240811_v3_final.pdf,

Accessed on 06/12/2016.

- Keys, C. W., Hand, B., Prain, V., & Collins, S. (1999). Using the science writing heuristic as a tool for learning from laboratory investigations in secondary science. *Journal of Research in Science Teaching*, **36**(10), 1065-1084.
- Klein, G. C., & Carney, J. M. (2014). Comprehensive approach to the development of communication and critical thinking: Bookend courses for third- and fourth-year chemistry majors. *Journal of Chemical Education*, **91**, 1649-1654.
- Kline, T. (2005). *Psychological testing : A practical approach to design and evaluation*. Thousand Oaks, Calif.: Thousand Oaks, Calif. : Sage Publications.
- Kogut, L. S. (1993). Critical thinking in general chemistry. *Journal of Chemical Education*, **73**(3), 218-221.
- Krejcie, R. V., & Morgan, D. W. (1970). Determining sample size for research activities. *Educational and Psychological Measurement*, **30**(3), 607-610.
- Kuhn, D. (1999). A developmental model of critical thinking. *Educational Researcher*, **28**(2), 16-26.
- Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Science*, **9**(5), 178-181.
- Lawson, A. E. (2000). *Classroom test of scientific reasoning. Multiple choice version*. Retrieved from <http://www.public.asu.edu/~anton1/AssessArticles/Assessments/Mathematics%20Assessments/Scientific%20Reasoning%20Test.pdf>, Accessed on 05/07/2015.
- Leggett, M., Kinnear, A., Boyce, M., & Bennett, I. (2004). Student and staff perceptions of the importance of generic skills in science. *Higher Education Research & Development*, **23**, 295-312.
- Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist*, **43**, 431-442.

- Lehman, D. R., & Nisbett, R. E. (1990). A longitudinal study of the effects of undergraduate training on reasoning. *Developmental Psychology*, **26**, 952-960.
- Lindsay, E. (2015). *Graduate outlook 2014 employers' perspectives on graduate recruitment in Australia*. Melbourne: Graduate Careers Australia. Retrieved from http://www.graduatecareers.com.au/wp-content/uploads/2015/06/Graduate_Outlook_2014.pdf, Accessed on 21/08/2015.
- Loo, S. R., & Thorpe, K. (1999). A psychometric investigation of scores on the Watson-Glaser critical thinking appraisal new form s. *Educational and Psychological Measurement*, **59**(6), 995-1003.
- Lowden, K., Hall, S., Elliot, D., & Lewin, J. (2011). *Employers' perceptions of the employability skills of new graduates: Research commissioned by the edge foundation*. Retrieved from http://www.educationandemployers.org/wp-content/uploads/2014/06/employability_skills_as_pdf_-_final_online_version.pdf, Accessed on 06/12/2016.
- Macpherson, K., & Owen, C. (2010). Assessment of critical thinking ability in medical students. *Assessment & Evaluation in Higher Education*, **35**(1), 41-54.
- Mann, L. (1979). *On the trail of process : A historical perspective on cognitive processes and their training*. New York: Grune & Stratton.
- Martineau, E., & Boisvert, L. (2011). Using wikipedia to develop students' critical analysis skills in the undergraduate chemistry curriculum. *Journal of Chemical Education*, **88**, 769-771.
- McMillan, J. (1987). Enhancing college students' critical thinking: A review of studies. *Journal of the Association for Institutional Research*, **26**(1), 3-29.
- McPeak, J. E. (1981). *Critical thinking and education*. Oxford: Martin Roberston.
- McPeak, J. E. (1990). *Teaching critical thinking : Dialogue and dialectic*. New York: New York : Routledge.

- Monash University. (2015). *Undergraduate - area of study. Chemistry*. Retrieved from <http://www.monash.edu.au/pubs/2015handbooks/aos/chemistry/>, Accessed on 15/04/2015.
- Moore, T. J. (2011). Critical thinking and disciplinary thinking: A continuing debate. *Higher Education Research & Development*, **30**(3), 261-274.
- Moore, T. J. (2013). Critical thinking: Seven definitions in search of a concept. *Studies in Higher Education*, **38**(4), 506-522.
- Nisbett, R. E., Fong, G. T., Lehman, D. R., & Cheng, P. W. (1987). Teaching reasoning. *Science (New York, N.Y.)*, **238**, 625-631.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: New York : McGraw-Hill.
- O'Hare, L., & McGuinness, C. (2015). The validity of critical thinking tests for predicting degree performance: A longitudinal study. *International Journal of Educational Research*, **72**, 162-172.
- OECD. (2014). *Pisa 2012 results: Creative problem solving: Students' skills in tackling real-life problems (volume v)*. OECD Publishing. Retrieved from <http://dx.doi.org/10.1787/9789264208070-en>, Accessed on 05/01/2018.
- Oliver-Hoyo, M. T. (2003). Designing a written assignment to promote the use of critical thinking skills in an introductory chemistry course. *Journal of Chemical Education*, **80**, 899-903.
- Ontario University. (2017). *Appendix 1: OCAV's undergraduate and graduate degree level expectations*. Retrieved from <http://oucqa.ca/framework/appendix-1/>, Accessed on 09/10/2017.
- Overton, T., Potter, N., & Leng, C. (2013). A study of approaches to solving open-ended problems in chemistry. *Chemistry Education Research and Practice*, **14**(4), 468-475.
- Pallant, J. F. (2016). *SPSS survival manual* (6th ed.). Sydney: Allen & Unwin.

- Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the hospital anxiety and depression scale (hads). *British Journal of Clinical Psychology*, **46**(1), 1-18.
- Pascarella, E. (1999). The development of critical thinking: Does college make a difference? *Journal of College Student Development*, **40**(5), 562-569.
- Pearson. (2015). *Watson-Glaser critical thinking appraisal - short form (WGCTA-S)*. Retrieved from <https://www.pearsonclinical.com.au/products/view/208>, Accessed on 03/07/2015.
- Phillips, V., & Bond, C. (2004). Undergraduates' experiences of critical thinking. *Higher Education Research & Development*, **23**(3), 277-294.
- Pithers, R. T., & Soden, R. (1999). Assessing vocational tutors' thinking skills. *Journal of Vocational Education & Training*, **51**(1), 23-37.
- Pithers, R. T., & Soden, R. (2000). Critical thinking in education: A review. *Educational Research*, **42**(3), 237-249.
- Prinsley, R., & Baranyai, K. (2015). *STEM skills in the workforce: What do employers want?* Retrieved from http://www.chiefscientist.gov.au/wp-content/uploads/OPS09_02Mar2015_Web.pdf, Accessed on 06/10/2015.
- Quitadamo, I. J., Kurtz, M. J., Cornell, C. N., Griffith, L., Hancock, J., & Egbert, B. (2011). Critical-thinking grudge match: Biology vs. Chemistry--examining factors that affect thinking skill in nonmajors science.(report). *Journal of College Science Teaching*, **40**(3), 19-25.
- Ramirez, T. V. (2017). On pedagogy of personality assessment: Application of Bloom's taxonomy of educational objectives. *Journal of Personality Assessment*, **99**(2), 146-152.
- Ramsden, P. (1992). *Learning to teach in higher education*. New York, NY: Routledge.
- Randles, C. A., & Overton, T. L. (2015). Expert vs. Novice: Approaches used by chemists when solving open-ended problems. *Chemistry Education Research and Practice*, **16**(4), 811-823.

- Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, **138**(2), 353-387.
- Sarkar, M., Overton, T., Thompson, C., & Rayner, G. (2016). Graduate employability: Views of recent science graduates and employers. *International Journal of Innovation in Science and Mathematics Education*, **24**(3), 31-48.
- Schreiber, L. M., & Valle, B. E. (2013). Social constructivist teaching strategies in the small group classroom. *Small Group Research*, **44**(4), 395-411.
- Stein, B., & Haynes, A. (2011). Engaging faculty in the assessment and improvement of students' critical thinking using the critical thinking assessment test. *Change: The Magazine of Higher Learning*, **43**(2), 44-49.
- Stein, B., Haynes, A., Redding, M., Ennis, T., & Cecil, M. (2007). Assessing critical thinking in STEM and beyond. In M. Iskander (Ed.), *Innovations in e-learning, instruction technology, assessment, and engineering education* (pp. 79-82): Springer Netherlands.
- Stephenson, N. S., & Sadler-Mcknight, N. P. (2016). Developing critical thinking skills using the science writing heuristic in the chemistry laboratory. *Chemistry Education Research and Practice*, **17**(1), 72-79.
- TalentLens. (2011). *Watson-Glaser critical thinking appraisal user-guide and technical manual UK supervised and unsupervised versions 2012*. UK: Pearson. Retrieved from <http://talentlens.com.sg/wp-content/uploads/downloads/2015/06/04-Watson-Glaser-user-guide-and-technical-manual-compressed.pdf>, Accessed on 08/12/2015.
- TalentLens. (2017). *The gold standard critical thinking test*. Retrieved from <http://www.thinkwatson.com/assessments/watson-glaser>, Accessed on
- Terenzini, P., Springer, L., Pascarella, E., & Nora, A. (1995). Influences affecting the development of students' critical thinking skills. *Journal of the Association for Institutional Research*, **36**(1), 23-39.

- The Critical Thinking Co. (2017). *Cornell critical thinking tests*. Retrieved from <https://www.criticalthinking.com/cornell-critical-thinking-tests.html>, Accessed on 9/10/2017.
- Thorndike, E. L., & Woodworth, R. S. (1901a). The influence of improvement in one mental function upon the efficiency of other functions. (i). *Psychological Review*, **8**(3), 247-261.
- Thorndike, E. L., & Woodworth, R. S. (1901b). The influence of improvement in one mental function upon the efficiency of other functions. ii. The estimation of magnitudes. *Psychological Review*, **8**(4), 384-395.
- Thorndike, E. L., & Woodworth, R. S. (1901c). The influence of improvement in one mental function upon the efficiency of other functions: Functions involving attention, observation and discrimination. *Psychological Review*, **8**(6), 553-564.
- Tiruneh, D. T., De Cock, M., Weldeslassie, A. G., Elen, J., & Janssen, R. (2016). Measuring critical thinking in physics: Development and validation of a critical thinking test in electricity and magnetism. *International Journal of Science and Mathematics Education*, 1-20.
- Tiruneh, D. T., Verburch, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. *Higher Education Studies*, **4**(1), 1-17.
- Tsai, C. C. (2001). A review and discussion of epistemological commitments, metacognition, and critical thinking with suggestions on their enhancement in internet-assisted chemistry classrooms. *Journal of Chemical Education*, **78**(7), 970-974.
- Tsui, L. (2002). Fostering critical thinking through effective pedagogy: Evidence from four institutional case studies. *The Journal of Higher Education*, **73**(6), 740-763.
- University of Adelaide. (2015). *University of Adelaide graduate attributes*. Retrieved from <http://www.adelaide.edu.au/learning/strategy/gradattributes/>, Accessed on 15/04/2015.

University of Edinburgh. (2017). *The University of Edinburgh's graduate attributes*. Retrieved from <http://www.ed.ac.uk/employability/graduate-attributes/framework>, Accessed on 09/10/2017.

University of Melbourne. (2015). *Handbook - chemistry*. Retrieved from <https://handbook.unimelb.edu.au/view/2015/!R01-AA-MAJ%2B1007>, Accessed on 15/04/2015.

Wason, P. C. (1966). New horizons. In B. Foss (Ed.), *Psychology*. Harmondsworth, England: Penguin.

Watson, G., & Glaser, E. M. (2006). *Watson-Glaser critical thinking appraisal short form manual*. San Antonio, TX: Pearson.

Weaver, M. G., Samoshin, A., Lewis, R., & Gainer, M. (2016). Developing students' critical thinking, problem solving, and analysis skills in an inquiry-based synthetic organic laboratory course. *Journal of Chemical Education*, **93**(5), 847-851.

Chapter 6 Appendices

Appendix A: Monash University Human Research Ethics Committee Certificate for Qualitative Research Investing Student's, Teaching Staff and Employer's Understanding of Critical Thinking



MONASH University

Monash University Human Research Ethics Committee (MUHREC)
Research Office

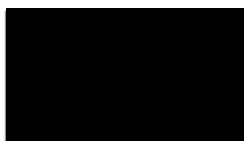
Human Ethics Certificate of Approval

This is to certify that the project below was considered by the Monash University Human Research Ethics Committee. The Committee was satisfied that the proposal meets the requirements of the *National Statement on Ethical Conduct in Human Research* and has granted approval.

Project Number: CF15/560 - 2015000258
Project Title: Perspectives of Skill Development at University
Chief Investigator: Dr Chris Thompson
Approved: From: 23 February 2015 To: 23 February 2020

Terms of approval - Failure to comply with the terms below is in breach of your approval and the Australian Code for the Responsible Conduct of Research.

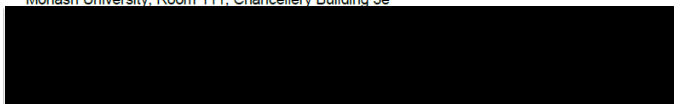
1. The Chief investigator is responsible for ensuring that permission letters are obtained, if relevant, before any data collection can occur at the specified organisation.
2. Approval is only valid whilst you hold a position at Monash University.
3. It is the responsibility of the Chief Investigator to ensure that all investigators are aware of the terms of approval and to ensure the project is conducted as approved by MUHREC.
4. You should notify MUHREC immediately of any serious or unexpected adverse effects on participants or unforeseen events affecting the ethical acceptability of the project.
5. The Explanatory Statement must be on Monash University letterhead and the Monash University complaints clause must include your project number.
6. **Amendments to the approved project (including changes in personnel):** Require the submission of a Request for Amendment form to MUHREC and must not begin without written approval from MUHREC. Substantial variations may require a new application.
7. **Future correspondence:** Please quote the project number and project title above in any further correspondence.
8. **Annual reports:** Continued approval of this project is dependent on the submission of an Annual Report. This is determined by the date of your letter of approval.
9. **Final report:** A Final Report should be provided at the conclusion of the project. MUHREC should be notified if the project is discontinued before the expected date of completion.
10. **Monitoring:** Projects may be subject to an audit or any other form of monitoring by MUHREC at any time.
11. **Retention and storage of data:** The Chief Investigator is responsible for the storage and retention of original data pertaining to a project for a minimum period of five years.



Professor Nip Thomson
Chair, MUHREC

cc: Prof Tina Overton, Mr Stephen Danczak

Monash University, Room 111, Chancellery Building 3e



Appendix B: Monash University Human Research Ethics Committee

Certificate for Research of the Danczak-Overton-Thompson Chemistry

Critical Thinking Test (DOT-CCTT)



Human Ethics Certificate of Approval

This is to certify that the project below was considered by the Monash University Human Research Ethics Committee. The Committee was satisfied that the proposal meets the requirements of the *National Statement on Ethical Conduct in Human Research* and has granted approval.

Project Number: CF16/568 - 2016000279

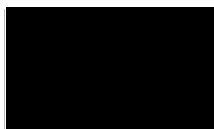
Project Title: Evaluating chemistry undergraduates' critical thinking skills

Chief Investigator: Dr Christopher Thompson

Approved: From: 26 February 2016 To: 26 February 2021

Terms of approval - Failure to comply with the terms below is in breach of your approval and the Australian Code for the Responsible Conduct of Research.


1. The Chief investigator is responsible for ensuring that permission letters are obtained, if relevant, before any data collection can occur at the specified organisation.
2. Approval is only valid whilst you hold a position at Monash University.
3. It is the responsibility of the Chief Investigator to ensure that all investigators are aware of the terms of approval and to ensure the project is conducted as approved by MUHREC.
4. You should notify MUHREC immediately of any serious or unexpected adverse effects on participants or unforeseen events affecting the ethical acceptability of the project.
5. The Explanatory Statement must be on Monash University letterhead and the Monash University complaints clause must include your project number.
6. **Amendments to the approved project (including changes in personnel):** Require the submission of a Request for Amendment form to MUHREC and must not begin without written approval from MUHREC. Substantial variations may require a new application.
7. **Future correspondence:** Please quote the project number and project title above in any further correspondence.
8. **Annual reports:** Continued approval of this project is dependent on the submission of an Annual Report. This is determined by the date of your letter of approval.
9. **Final report:** A Final Report should be provided at the conclusion of the project. MUHREC should be notified if the project is discontinued before the expected date of completion.
10. **Monitoring:** Projects may be subject to an audit or any other form of monitoring by MUHREC at any time.
11. **Retention and storage of data:** The Chief Investigator is responsible for the storage and retention of original data pertaining to a project for a minimum period of five years.



Professor Nip Thomson
Chair, MUHREC

cc: Mr Stephen Danczak, Prof Tina Overton, Dr Laurence Orlando

Monash University, Room 111, Chancellery Building E



Appendix C: Student Questionnaire Regarding Understanding of Critical Thinking

**Monash School of Chemistry:
Undergraduate skill development survey**

In order of importance (1 = most important), list the skills you believe are most important for you to develop while at university?

1. _____
2. _____
3. _____
4. _____
5. _____

PLEASE COMPLETE THIS PAGE BEFORE READING ANY FURTHER

On January 1st, 2015 what was your age

How many units have you completed in the School of Chemistry

With which sex do you identify?

☐ Female

☐ Male

☐ Other

☐ Rather not say

1. What does the term “Critical Thinking” mean to you?

2. Please consider Statements A and B:

Statement A:

Developing my critical thinking skills at University is important

Statement B:

Developing my critical thinking skills at University is not important

Now, please choose one of the options below which best matches your view on developing your critical thinking skills at University:

Strongly
Agree
with A

Agree
with A

Neutral

Agree
with B

Strongly Agree
with B

3. Please provide an example of when you have had the opportunity to develop your critical thinking while studying chemistry?

4. Please comment on how confident you felt about your ability to think critically during the task you described in Q.3 (above)?

5. Please provide an example of when you have had the opportunity to think critically outside of your university studies?

Appendix D: Student Questionnaire Regarding Understanding of Critical Thinking

**Monash School of Chemistry:
Teaching staff student skill development survey**

In order of importance (1 = most important), list the skills you believe are most important for students to develop while at university?

1. _____
2. _____
3. _____
4. _____
5. _____

PLEASE COMPLETE THIS PAGE BEFORE READING ANY FURTHER

Which of the following activities did you undertake as part of your teaching within the school of chemistry in 2014? *(choose any relevant boxes)*

☐Lecturing

☐Tutoring

☐Laboratory Teaching

Which year level(s) did you perform these activities? *(choose any relevant boxes)*

☐1st year

☐2nd year

☐3rd year

☐Honours

In total, how many units did you teach on within the school of chemistry, in some way in 2014?

1. What does the term “Critical Thinking” mean to you?

2. What is your view on the development of students’ critical thinking skills at University?

(choose one of: SA – Strongly Agree, A – Agree or N – Neutral with the statement that best matches your view)

Developing Students’
critical thinking skills at
University is important

SA A N A SA

Developing Students’
critical thinking skills at
University IS NOT
important

3. Can you provide an example of when you have provided students with the opportunity to develop their critical thinking while studying chemistry?

4. Can you comment on how confident you felt about your ability to teach students to think critically during the task you described in Q.3 (above)?

Appendix E: Danczak-Overton-Thompson Chemistry Critical Thinking

Test Version 1 (DOT-CCTTv1) with Answers

MAKING ASSUMPTIONS

Scientists often make assumptions when assessing the conclusions of other scientists' work or argument. Assumptions are statements for which no evidence or proof has been provided. Many assumptions are unstated.

For example:

A **valid assumption** is that when a chemical reaction was carried out it was done in clean glassware.

- It is taken for granted that scientists use clean equipment.

An **invalid assumption** would be that the reaction was carried out in the absence of air.

- Unless explicitly stated as to what conditions the reaction was carried out this cannot be taken for granted.

Instructions:

In this section you are presented with short statements followed by several assumptions.

- You need to decide if an assumption is a **(A) Valid Assumption** or an **(B) Invalid Assumption**.
- Base your choices **ONLY ON THE INFORMATION IN THE SHORT STATEMENTS**.
- Assume that **THE INFORMATION IN THE SHORT STATEMENTS ARE TRUE**.

The statement below is used for Questions 1 – 4.

A chemist tested a metal centred complex by placing it in a magnetic field. The complex was attracted to the magnetic field. From this result the chemist decided the complex had unpaired electrons and was therefore paramagnetic rather than diamagnetic.

1. The metal within the complex is responsible for determining how a complex interacts with a magnetic field. **There is no indication that to be able make an assumption at all as to what causes complex to interact.**

A) Valid Assumption

B) Invalid Assumption

2. Diamagnetic metals centred complexes do not have any unpaired electrons. **The paragraph suggests that if the complex has unpaired electrons it is paramagnetic. This means diamagnetic complexes likely cannot have unpaired electrons.**

A) Valid Assumption

B) Invalid Assumption

3. Complexes can only ever be paramagnetic or diamagnetic. **There is no mention of the statement that these are the only two options.**

A) Valid Assumption

B) Invalid Assumption

4. A paramagnetic metal centred complex cannot be a diamagnetic metal centred complex. **The paragraph states that a complex is paramagnetic rather than diamagnetic, suggesting a complex cannot be both.**

A) Valid Assumption

B) Invalid Assumption

The statement below is used for Questions 5 – 7.

Carbonate (CO_3^{2-}) can accept two hydrogen ions (H^+). This is an example of a diprotic base. A monoprotic base such as hydroxide (OH^-) can only accept one hydrogen ion (H^+).

5. Carbonate (CO_3^{2-}) accepts hydrogen ions (H^+) more readily than monoprotic bases. **There is no mention of how readily either of the bases accept hydrogen ions.**

A) Valid Assumption

B) Invalid Assumption

6. Hydrogen ions (H^+) have greater attraction towards bases with a more negative charge. **There is no mention of a relationship between the degree on negative charge and hydrogen ion attraction.**

A) Valid Assumption

B) Invalid Assumption

7. The larger negative charge of a base, the more hydrogen ions (H^+) it can accept. **This requires being able to connect the two statements regarding how many hydrogen ion a base can accept.**

A) Valid Assumption

B) Invalid Assumption

ANALYSING ARGUMENTS

Scientists develop arguments based on their results or critique the arguments of others work.

A **Valid Argument**:

- Is directly related to the statement or question
- Is of significant importance
- Is supported by evidence

An **Invalid Argument**:

- Is not directly related to the statement or question
- Is not of significant importance
- Is not supported by evidence

Instructions:

- In this section you are presented with short statements followed by several arguments.
- You need to decide if each argument is a **(A) Valid Argument** or an **(B) Invalid Argument**.
- Base your choices **ONLY ON THE INFORMATION IN THE SHORT STATEMENTS**.
- Assume that **THE INFORMATION IN THE SHORT STATEMENTS ARE TRUE**.
- Assume that **THE INFORMATION IN THE ARGUMENTS ARE TRUE**.

The statement below is used for Questions 8 – 10.

Zinc oxide (ZnO) protects against DNA damage from UV radiation. Should it be removed from all sunscreens to decrease exposure to toxins?

8. Yes. Zinc oxide particles are smaller than 100nm (0.000000001 millimetres) and are small enough to be absorbed by skin cells. **While it is true that zinc oxide could be absorbed into cells there is no discussion to suggest that zinc oxide is a toxin.**

A) Valid Argument

B) Invalid Argument

9. Yes. There are other chemicals which could be used in sunscreens to absorb UV radiation before it can do damage to DNA. **While there are other chemicals that could absorb UV radiation, the argument is whether removing zinc oxide will reduce exposure to toxins, which is not discussed.**

A) Valid Argument

B) Invalid Argument

10. No. Zinc oxide effectively reflects and diffracts harmful UV-radiation. Removing it would increase exposure to UV-radiation. **This argument does not discuss the toxicity of zinc oxide, however reducing the damage to DNA caused by UV radiation is of significant importance making this a valid argument.**

A) Valid Argument

B) Invalid Argument

The statement below is used for Questions 11 – 14.

Should it be illegal to produce polycarbonate bottles containing bisphenol A (BPA)?

11. Yes. The low-dose hypothesis states: health effects occur at doses far below previous levels. These previous levels were determined to be safe using well established toxicological procedures and principles. **As this is a hypothesis there is no evidence to support this claim despite the argument being significant.**

A) Valid Argument

B) Invalid Argument

12. Yes. Bisphenol A has a chemical structure which could interact with DNA. **This could be considered a significant argument however there is no evidence to suggest that an interaction will occur or whether an interaction is detrimental.**

A) Valid Argument

B) Invalid Argument

13. No. Government and academic researchers estimate the intake of BPA is less than 0.00000125 milligrams per kilogram of body weight per day. The acceptable dose of BPA put forth by the US Environmental Protection Agency is 0.05 milligrams per kilogram of body weight per day. **This argument addresses a significant point and is supported by what would be considered credible sources.**

A) Valid Argument

B) Invalid Argument

14. No. several independent studies from toxicology experts were unable to replicate the finding of the University of Missouri that low doses of BPA led to adverse effects in mice. **This evidence suggests that the claims made by the University of Missouri are not reliable and there is no reason to remove BPA based on the Universities results.**

A) Valid Argument

B) Invalid Argument

DEVELOPING HYPOTHESES

Scientists often collect observations to develop hypotheses. These are called inferences which are purely based on the observations presented to the researcher. They are never certain however some inferences are more accurate than others.

A hypothesis is **Likely to be an accurate inference** when:

- It draws conclusions only from the observations
- Assumptions are reasonable

Insufficient Information to determine accuracy of a hypothesis occurs when:

- More detailed observations are required
- Assumptions may be reasonable but require further evidence

A hypothesis is **Unlikely to be an accurate inference** when:

- The conclusions are drawn from outside of the observations
- Assumptions are unreasonable or require further evidence

Instructions:

- In this section you are presented with short statements followed by several inferences.
- You need to decide if each inference is/has:
 - A) **Likely to be an accurate inference**
 - B) **Insufficient Information to determine accuracy**
 - C) **Unlikely to be an accurate inference**

The passage below is used for Questions 15 – 17.

The following is an excerpt from an article posted by NASA at the URL: www.nasa.gov/press-release/nasa-confirms-evidence-that-liquid-water-flows-on-today-s-mars

Using an imaging spectrometer researchers detected signatures of hydrated minerals on slopes where mysterious streaks are seen on the Red Planet. These darkish streaks appear to ebb and flow over time. They darken and appear to flow down steep slopes during warm seasons, and then fade in cooler seasons. They appear in several locations on Mars when temperatures are above minus 10 degrees Fahrenheit (minus 23 Celsius), and disappear at colder times.

“Our quest on Mars has been to ‘follow the water,’ in our search for life in the universe, and now we have convincing science that validates what we’ve long suspected,” said John Grunsfeld, astronaut and associate administrator of NASA’s Science Mission Directorate in Washington. “This is a significant development, as it appears to confirm that water -- albeit briny -- is flowing today on the surface of Mars.”

These downhill flows, known as recurring slope lineae (RSL), often have been described as possibly related to liquid water. The new findings of hydrated salts on the slopes point to what that relationship may be to these dark features. The hydrated salts would lower the freezing point of a liquid brine, just as salt on roads here on Earth causes ice and snow to melt more rapidly. Scientists say it’s likely a shallow subsurface flow, with enough water wicking to the surface to explain the darkening.

15. It is a fact that water can exist in a liquid state below zero degrees Celsius due to the presence of the hydrated salts. On Mars this could lead to the freezing and thawing of water resulting in the ebb and flow of the mysterious streaks on Mars. **This analogy is discussed in the third paragraph and it would be reasonable to assume that this natural phenomena can occur on Earth and Mars.**

- A) Likely to be an accurate inference
- B) Insufficient Information to determine accuracy
- C) Likely to be inaccurate inference

16. Some planets are likely to have similar conditions to Mars and NASA is planning to determine if there may be habitable water on these other planets. **While the second paragraph mentions ‘in our search for life in the universe’ and it is reasonable to assume a group like NASA are actively pursuing this goal, the assumption is drawn from outside what is presented and more information would be required to decide if there were plans in place.**

- A) Likely to be an accurate inference
- B) Insufficient Information to determine accuracy
- C) Likely to be inaccurate inference

17. The imaging spectrometer directly measures water on Mars. **There was no mention that water was measured directly. This conclusion is drawn from outside what is presented. The first paragraph describes that hydrated minerals were detected via spectrometry not water itself.**

- A) Likely to be an accurate inference
- B) Insufficient Information to determine accuracy
- C) Likely to be inaccurate inference

The passage below is used for Questions 18 – 20.

The following is an excerpt from an article posted by CSIRO at the URL:
www.csiro.au/en/News/News-releases/2015/Marine-debris

Researchers from CSIRO and Imperial College London have assessed how widespread the threat of plastic is for the world's seabirds and found the majority of species have plastic in their gut.

The study, led by Dr Chris Wilcox with co-authors Dr Denise Hardesty and Dr Erik van Seville and published today in the journal PNAS, found that nearly 60 per cent of all seabird species have plastic in their gut. Based on analysis of published studies since the early 1960s, the researchers found that plastic is increasingly common in seabird's stomachs.

In 1960, plastic was found in the stomach of less than 5 per cent of individual seabirds, rising to 80 per cent by 2010. The researchers predict that plastic ingestion will affect 99 per cent of the world's seabird species by 2050, based on current trends. The scientists estimate that 90 per cent of all seabirds alive today have eaten plastic of some kind.

This includes bags, bottle caps, and plastic fibres from synthetic clothes, which have washed out into the ocean from urban rivers, sewers and waste deposits. Birds mistake the brightly coloured items for food, or swallow them by accident, and this causes gut impaction, weight loss and sometimes even death.

18. This passage implies that as a result of 90 per cent of all sea birds having eaten plastic of some kind, some species of sea bird become endangered. **The fourth paragraph states sometimes death occurs in birds which have consumed plastic however to suggest this would lead to some sea birds becoming endangered is a conclusion drawn from outside the passage.**

- A) Likely to be an accurate inference
- B) Insufficient Information to determine accuracy
- C) Likely to be inaccurate inference**

19. The number of seabird to have eaten plastic of some kind has a direct link to pollution caused consumer waste. **While it not certain there is explicit mention of plastics such as bags, bottle caps and clothing washing out to sea and it is reasonable to assume these are forms of consumer waste.**

- A) Likely to be an accurate inference**
- B) Insufficient Information to determine accuracy
- C) Likely to be inaccurate inference

20. The number of seabirds to have eaten plastic is responsible for a decline in biodiversity. **There is discussion in the final paragraph of birds mistaking brightly coloured plastics for food however to assume a causation between the two would require additional information.**

- A) Likely to be an accurate inference
- B) Insufficient Information to determine accuracy**
- C) Likely to be inaccurate inference

TESTING HYPOTHESES

Often Scientists conduct experiments to test a hypothesis. Based on what they observe they make deductions.

A deduction is based on cause and effect.

A **Logical Deduction** is when evidence is observed to support a hypothesis.

An **Illogical Deduction** is when there insufficient evidence to support a hypothesis.

Instructions:

- In this section you are presented with short statements followed by several hypotheses.
- You need to decide if each hypothesis is a **(A) Logical Deduction** or an **(B) Illogical Deduction**.
- Base your choice **ONLY ON THE INFORMATION IN THE SHORT STATEMENTS**.
- Assume that **THE INFORMATION IN THE SHORT STATEMENTS ARE TRUE**.

The statement below is used for Questions 21 – 23.

When making aspirin a chemist needed to make an ester bond. However, she only had access to a carboxylic acid. The carboxylic acid is less likely to form an ester bond than if an anhydride were used.

21. The formation of aspirin will produce a low yield. **The passage refers to likelihood of formation of aspirin and this cannot necessarily be related to yield.**

A) Logical Deduction

B) Illogical Deduction

22. The formation of aspirin using carboxylic acid is more likely to fail than if an anhydride were used. **The passage clearly states that a carboxylic acid will be less successful than an anhydride.**

A) Logical Deduction

B) Illogical Deduction

23. Anhydrides are more likely to succeed in forming an ester bond than carboxylic acids. **This is a rewording of the previous question also looking to identify that the passage clearly states that a carboxylic acid will be less successful than an anhydride.**

A) Logical Deduction

B) Illogical Deduction

The statement below is used for Questions 24 and 25.

The most recent isolation of Taxol from the plant *Taxus brevifolia* had the highest percentage yield as compared to the previous six trials. The percentage yield calculations have a 5% uncertainty associated with them.

24. The percentage yield of Taxol was higher than expected. **There is no mention of the expectation.**

A) Logical Deduction

B) Illogical Deduction

25. The percentage yield of Taxol in the most recent trial was greater than first trial. **The passage states that yield was the highest of the previous six trials.**

A) Logical Deduction

B) Illogical Deduction

DRAWING CONCLUSION

Often scientists will be presented with a series of observation, data or information. They are required to arrive a conclusions based on several pieces of information.

A **Reasonable Conclusion** is:

- Able to demonstrate the possible effect caused by the observations made
- Is likely to be true but may not be absolutely accurate

An **Unreasonable Conclusion** is:

- Unable to demonstrate the possible effect caused by the observations made
- Is likely to be untrue

Instructions:

- In this section you are presented with short statements followed by several hypotheses.
- You need to decide if each conclusion is a (A) Reasonable Conclusion or an (B) Unreasonable Conclusion.
- Base your choices ONLY ON THE INFORMATION IN THE SHORT STATEMENTS.
- Assume that THE INFORMATION IN THE SHORT STATEMENTS ARE TRUE.

The reactivity of a carbon centred molecule is related to the stability of its intermediate. The stability of the intermediate is dependent on the number of carbon atoms attached to the carbon centre. Adjacent carbon atoms can distribute the charge on the intermediate.

A) Reasonable Conclusion B) Unreasonable Conclusion

A) Reasonable Conclusion B) Unreasonable Conclusion

At rest a typical cell membrane has a concentration of 140 mM (millimolar) of potassium ions (K^+) within the cell. The concentration of potassium ions (K^+) within the cell can be decreased via open voltage gated potassium channels. The voltage gated potassium channels will only open if the concentration of sodium ions (Na^+) is greater than 5 mM within the cell. At rest a typical cell membrane has concentration 5 mM of sodium ions (Na^+) within the cell.

A) Reasonable Conclusion B) Unreasonable Conclusion

A) Reasonable Conclusion B) Unreasonable Conclusion

A) Reasonable Conclusion B) Unreasonable Conclusion

Appendix F: DOT-CCTTv1 Statistics Summary

Table F1 Summary of descriptive statistics for DOT-CCTTv1 score

		DOT-CCTTv1 Score
n		615
Mean		17.55
95% Confidence Interval for Mean	Lower Bound	17.28
	Upper Bound	17.82
Median		18.00
Variance		11.417
Standard Deviation		3.379
Minimum		7
Maximum		30
Range		23
Interquartile Range		5
Skewness		.034
Kurtosis		-.274

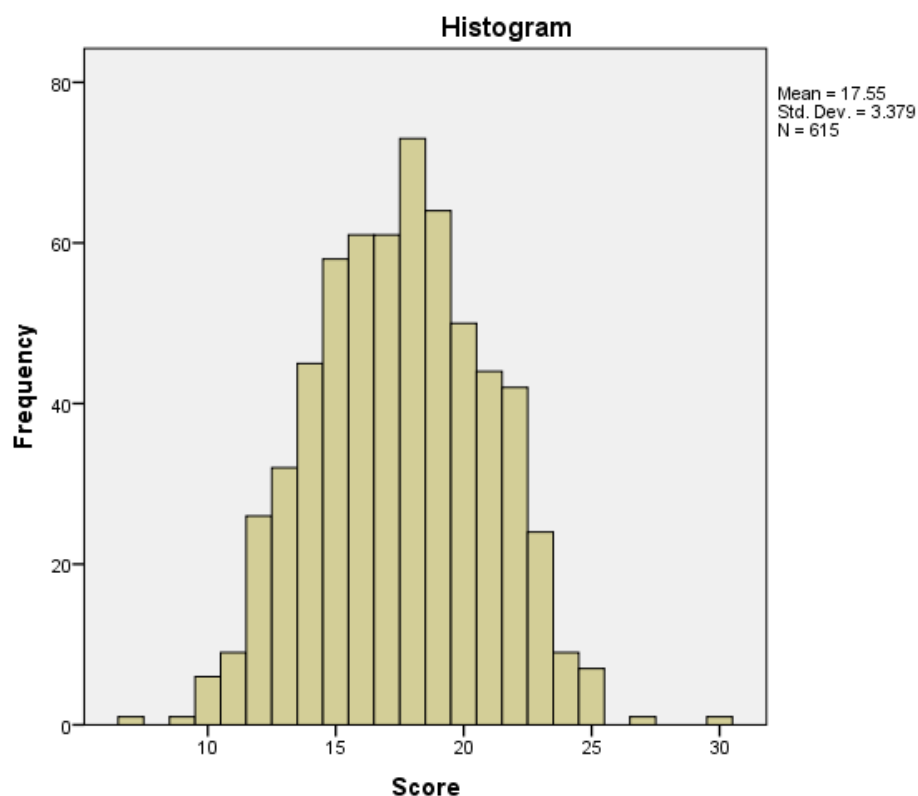


Figure F1 Histogram output from IBM SPSS of DOT-CCTTv1 scores versus frequency

Table F2 Scale reliability of DOT-CCTTv1 score treating the DOT-CCTTv1 as a single scale of 30 items

Cronbach's α	Number of Items including scale	Cases	Number of Cases	Percentage of Cases (%)
.625	31	Valid	587	95.4
		Excluded ^a	28	4.6
		Total	615	100.0

^aListwise (cases with data missing from the variable(s) of interest) deletion based on all variables in the procedure.

Table F3 Scale reliability of DOT-CCTTv1 score treating the DOT-CCTTv1 as single scale made of five sub-scales

Cronbach's α	Number of Items including scale	Cases	Number of Cases	Percentage of Cases (%)
.682	6	Valid	615	100.0
		Excluded ^a	0	0.0
		Total	615	100.0

^aListwise (cases with data missing from the variable(s) of interest) deletion based on all variables in the procedure.

Table F4 Summary of correct item total correlations and Cronbach's α if item deleted of the five sub-scales of the DOT-CCTTv1 score

Sub-scale	Corrected Item-Total Correlation	Cronbach's α if Item Deleted
Making Assumptions	.412	.644
Analysing Arguments	.368	.658
Developing Hypotheses	.421	.643
Testing Hypotheses	.250	.685
Drawing Conclusions	.369	.662

Table F5 Example reliability of sub-scale score 'Making Assumptions'

Cronbach's α	Number of Items including sub-scale	Cases	Number of Cases	Percentage of Cases (%)
.648	8	Valid	609	99.0
		Excluded ^a	6	1.0
		Total	615	100.0

^aListwise (cases with data missing from the variable(s) of interest) deletion based on all variables in the procedure

Table F6 Summary of correct item total correlations and Cronbach's α if item deleted of questions in the sub-scale 'Making Assumptions'

Questions	Corrected Item-Total Correlation	Cronbach's α if Item Deleted
Q1	.250	.638
Q2	.312	.626
Q3	.205	.646
Q4	.242	.639
Q5	.392	.611
Q6	.396	.611
Q7	.288	.632

Table F7 Summary of Split Halves reliability for the DOT-CCTTv1

Cronbach's α	Part 1	α	.291
		Number of Items	15 ^a
	Part 2	α	.225
		Number of Items	15 ^b
	Total Number of Items		
Correlation Between Part 1 and Part 2			.304
Spearman-Brown Coefficient (ρ_{cc})			.467

^a. The items are: Q1, Q3, Q5, Q7, Q9, Q11, Q13, Q15, Q17, Q19, Q21, Q23, Q25, Q27, Q29.

^b. The items are: Q2, Q4, Q6, Q8, Q10, Q12, Q14, Q16, Q18, Q20, Q22, Q24, Q26, Q28, Q30.

Table F8 Sample Mann-Whitney U comparing median scores of participants who answered question 1 of the sub-scale 'Making Assumption' correctly and those who answered it incorrectly

Question 1	Correct	Number of Cases	196
		Median	19
	Incorrect	Number of Cases	413
		Median	17
	Total number of Cases		609
Mann-Whitney U			16129.00
Standardised Test Statistic (z score)			9.280
Significance (p)			.000
Effect size (r)			.24

Appendix G: DOT-CCTTv1 Sample Academic Transcript and Coding

Table G Codes and corresponding themes used in qualitative analysis of academic transcript

Codes	Themes
IC	Instruction Clarity
WQ	Wording of the Question
IPS	Information within the Parent Statement
PK	Prior Knowledge

- P4: Your opening statement is unclear. **[WQ/IPS]**
- P6: So do you expect students to provide an explanation?
- I: No.
- P6: No?
- I: I mean like as a concession this is not the best way to go about assessing someone's critical thinking on a one to one basis but if you've got a thousand students it's kind of the best option you've got. Yeah I mean you would need a lot of rigorous staff to do open ended stuff which is probably not feasible on a large cohort level.
- I: Um, question 10. Valid or a valid argument?
- P6: Valid.
- P5: Yeah.
- P6: Based on the information is valid. **[IPS]**
- P5: Yep.
- I: Invalid argument?
- P11: I'm gonna go for invalid – ah- because, the premise in the question was that it should be removed to decrease exposure to toxins. **[IPS]**
- I: Mhmm
- P11: However the conclusion of number ten saying removing it would increase exposure to UV radiation which in itself is not a toxin. **[IPS]**
- I: Mhmm
- AP: So they're not logically connected but again it's a desirable outcome. **[IPS/IC]**
- I: Mhmm
- P11: So I don't believe they're necessarily connected. **[IPS/IC]**
- P1: So it's a good point but not a valid argument? **[IPS/IC]**
- P11: Well yeeees....
- P3: Well it is valid 'cause we're talking about its role in sun screen which is in the role of been out in the sun. **[IPS/PK]**
- P11: Yeah, yeah, yeah, yeah. But the premise of the statement, sorry the um-question posed was should it be removed to decrease exposure to toxins. Nothing about should it be removed and therefore replaced to prevent sun exposure. So it – cause I see where it comes from but it doesn't follow from the exposure to toxins. **[IPS/PK]**
- P4: It's a leap of faith which I don't where science leads. **[IPS/PK]**

Appendix H: Danczak-Overton-Thompson Chemistry Critical Thinking

Test Version 2 (DOT-CCTTv2) with Answers

MAKING ASSUMPTIONS

An assumption is something that's taken for granted. There is no additional proof or evidence required in the statement being made. Assumptions can be stated or implied. Scientists generally clarify their assumption in academic discussion however some assumptions are implied.

For example:

When reading the synthesis for a chemical procedure there is no mention of using clean glassware. It is generally assumed that the scientist adhered to standard cleaning methods and therefore this is a **VALID ASSUMPTION**.

If a synthesis stated that it used 98% ethanol it would be an **INVALID ASSUMPTION** to assume that the synthesis was free from impurities. In this case all the information to determine the validity of the assumption has been presented.

Instructions:

- In this section you are presented with short statements followed by several assumptions.
- Select **(A) Valid Assumption** if you think the assumption can be taken for granted based on the information presented in the short statement.
- Select **(B) Invalid Assumption** if you think the assumption *cannot* be taken for granted based on the information presented in the short statement.
- *Treat each question individually* and base your decisions relating only to the original statement.

The statement below is used for Questions 1 – 4.

Paramagnetic and diamagnetic metal complexes behave differently when exposed to a magnetic field. A chemist tested a metal complex by placing it in a magnetic field. From the result of the test the chemist decided the metal complex had unpaired electrons and was therefore paramagnetic.

1. The metal within the complex is responsible for determining how a metal complex interacts with a magnetic field. **There is no indication that to be able make an assumption at all as to what causes complex to interact.**

A) Valid Assumption

B) Invalid Assumption

2. Diamagnetic metal complexes do not have any unpaired electrons. **The paragraph suggests that if the complex has unpaired electrons it is paramagnetic. This means diamagnetic complexes likely cannot have unpaired electrons.**

A) Valid Assumption

B) Invalid Assumption

3. Metal complexes can only ever be paramagnetic or diamagnetic. **There is no mention of the statement that these are the only two options.**

A) Valid Assumption

B) Invalid Assumption

4. A paramagnetic metal complex cannot be a diamagnetic metal complex. **The paragraph states that a complex is paramagnetic rather than diamagnetic, suggesting a complex cannot be both.**

A) Valid Assumption

B) Invalid Assumption

The statement below is used for Questions 5 – 7.

Carbonate (CO_3^{2-}) has a formal charge of negative 2. Carbonate (CO_3^{2-}) can accept two hydrogen ions (H^+) which each a formal charge of positive 1. Carbonate (CO_3^{2-}) is an example of a diprotic base. A monoprotic base such as hydroxide (OH^-) has a formal charge of negative 1. Hydroxide (OH^-) can only accept one hydrogen ion (H^+).

5. Carbonate (CO_3^{2-}) accepts hydrogen ions (H^+) more easily than monoprotic bases. **There is no mention of how readily either of the bases accept hydrogen ions.**

A) Valid Assumption

B) Invalid Assumption

6. Hydrogen ions (H^+) have greater attraction towards bases with a more negative charge. **There is no mention of a relationship between the degree on negative charge and hydrogen ion attraction.**

A) Valid Assumption

B) Invalid Assumption

7. The greater the formal negative charge of a base, the more hydrogen ions (H^+) it can accept. **This requires relating the two statements regarding how many hydrogen ion a base can accept.**

A) Valid Assumption

B) Invalid Assumption

DEVELOPING HYPOTHESES

When generating hypotheses, scientists will draw inferences based on the information observed and the supposed facts. An inference is used to fill in the gaps to create a connection or look for the intended meaning. These inferences are not a certainty but based on the information available there is great deal of confidence in the hypothesis being developed.

For Example:

A chemist added one chemical to another and this was followed immediately by a colour change. It is **LIKELY TO BE AN ACCURATE INFERENCE** that this occurred due to a reaction between the two chemicals. There is the possibility that one of chemicals may have reacted with air or broken down to result in the colour change but *based on the observations* and the *reasonable assumption* that the chemicals will not break down or react with air this is **LIKELY TO BE AN ACCURATE INFERENCE**.

In the above example there would be **INSUFFICIENT INFORMATION TO DETERMINE ACCURACY** of a hypothesis which suggests one of the chemicals is more reactive than the other. *More detailed observation are required*, for example reacting the chemicals in question with other chemicals. Furthermore any *assumptions would require additional evidence*, for example, where the chemicals at the same concentration?

If a chemist were to suggest that when two chemicals react they will produce a colour change, this would be **UNLIKELY TO BE AN ACCURATE INFERENCE**. The *conclusion is drawn from outside the information* presented: they haven't tested other chemicals, and the make an *unreasonable assumption* that reaction result in a colour change.

Instructions:

- In this section you are presented with short statements followed by several inferences.
- You need to decide if each inference is/has:
 - A) **Likely to be an accurate inference**
 - B) **Insufficient Information to determine accuracy**
 - C) **Unlikely to be an accurate inference**

The passage below is used for Questions 8 – 10.

The following is an excerpt from an article posted by NASA at the URL: www.nasa.gov/press-release/nasa-confirms-evidence-that-liquid-water-flows-on-today-s-mars

Using an imaging spectrometer researchers detected signatures of hydrated minerals on slopes where mysterious streaks are seen on the Red Planet. These darkish streaks appear to ebb and flow over time. They darken and appear to flow down steep slopes during warm seasons, and then fade in cooler seasons. They appear in several locations on Mars when temperatures are above minus 10 degrees Fahrenheit (minus 23 Celsius), and disappear at colder times.

“Our quest on Mars has been to ‘follow the water,’ in our search for life in the universe, and now we have convincing science that validates what we’ve long suspected,” said John Grunsfeld, astronaut and associate administrator of NASA’s Science Mission Directorate in Washington. “This is a significant development, as it appears to confirm that water -- albeit briny -- is flowing today on the surface of Mars.”

These downhill flows, known as recurring slope lineae (RSL), often have been described as possibly related to liquid water. The new findings of hydrated salts on the slopes point to what that relationship may be to these dark features. The hydrated salts would lower the freezing point of a liquid brine, just as salt on roads here on Earth causes ice and snow to melt more rapidly. Scientists say it’s likely a shallow subsurface flow, with enough water wicking to the surface to explain the darkening.

8. It is a fact that water can exist in a liquid state below zero degrees Celsius due to the presence of the hydrated salts. On Mars this could lead to the freezing and thawing of water resulting in the ebb and flow of the mysterious streaks on Mars. **This analogy is discussed in the third paragraph and it would be reasonable to assume that this natural phenomena can occur on Earth and Mars.**

- A) Likely to be an accurate inference
- B) Insufficient Information to determine accuracy
- C) Likely to be inaccurate inference

9. Some planets are likely to have similar conditions to Mars and NASA is planning to determine if there may be habitable water on these other planets. **While the second paragraph mentions ‘in our search for life in the universe’ and it is reasonable to assume a group like NASA are actively pursuing this goal, the assumption is drawn from outside what is presented and more information would be required to decide if there were plans in place.**

- A) Likely to be an accurate inference
- B) Insufficient Information to determine accuracy
- C) Likely to be inaccurate inference

10. The imaging spectrometer directly measures water on Mars. **There was no mention that water was measured directly. This conclusion is drawn from outside what is presented. The first paragraph describes that hydrated minerals were detected via spectrometry not water itself.**

- A) Likely to be an accurate inference
- B) Insufficient Information to determine accuracy
- C) Likely to be inaccurate inference

The passage below is used for Questions 11 – 13.

The following is an excerpt from an article posted by CSIRO at the URL:

www.csiro.au/en/News/News-releases/2015/Marine-debris

Researchers from CSIRO and Imperial College London have assessed how widespread the threat of plastic is for the world's seabirds and found the majority of species have plastic in their gut.

The study, led by Dr Chris Wilcox with co-authors Dr Denise Hardesty and Dr Erik van Seville and published today in the journal PNAS, found that nearly 60 per cent of all seabird species have plastic in their gut. Based on analysis of published studies since the early 1960s, the researchers found that plastic is increasingly common in seabird's stomachs.

In 1960, plastic was found in the stomach of less than 5 per cent of individual seabirds, rising to 80 per cent by 2010. The researchers predict that plastic ingestion will affect 99 per cent of the world's seabird species by 2050, based on current trends. The scientists estimate that 90 per cent of all seabirds alive today have eaten plastic of some kind.

This includes bags, bottle caps, and plastic fibres from synthetic clothes, which have washed out into the ocean from urban rivers, sewers and waste deposits. Birds mistake the brightly coloured items for food, or swallow them by accident, and this causes gut impaction, weight loss and sometimes even death.

11. This passage implies that as a result of 90 per cent of all sea birds having eaten plastic of some kind, some species of sea bird become endangered. **The fourth paragraph states sometimes death occurs in birds which have consumed plastic however to suggest this would lead to some sea birds becoming endangered is a conclusion drawn from outside the passage.**

- A) Likely to be an accurate inference
- B) Insufficient Information to determine accuracy
- C) Likely to be inaccurate inference**

12. The number of seabird to have eaten plastic of some kind has a direct link to pollution caused consumer waste. **While it not certain there is explicit mention of plastics such as bags, bottle caps and clothing washing out to sea and it is reasonable to assume these are forms of consumer waste.**

- A) Likely to be an accurate inference**
- B) Insufficient Information to determine accuracy
- C) Likely to be inaccurate inference

13. The number of seabirds to have eaten plastic is responsible for a decline in biodiversity. **There is discussion in the final paragraph of birds mistaking brightly coloured plastics for food however to assume a causation between the two would require additional information.**

- A) Likely to be an accurate inference
- B) Insufficient Information to determine accuracy**
- C) Likely to be inaccurate inference

TESTING HYPOTHESES

Scientists conduct experiments to test a hypothesis. A deduction is drawn purely from the observations. They begin with a hypothesis or statement they believe to be true and systematically seek information to confirm or refute the hypothesis. This results in a premise believed to be accurate or true.

For Example:

After completing a synthesis, a chemist compares the product to the properties of the starting material. The hypothesis is that if the properties of the product are different that of the starting material there is no starting material in the product. They find that the product has very different properties to the starting material.

A **REASONABLE DEDUCTION** is that there is no starting material in the product. The *evidence presented is supported by this hypothesis*.

An **UNREASONABLE DEDUCTION** would be that based on these observation the product is the chemical the chemist was seeking to produce. *The deduction is not related to the hypothesis* and there is *insufficient evidence to support this claim*.

Instructions:

- In this section you are presented with short statements followed by several hypotheses.
- You need to decide if each hypothesis is a **(A) REASONABLE DEDUCTION** or an **(B) UNREASONABLE DEDUCTION**.
- Base your choice **ONLY ON THE INFORMATION IN THE SHORT STATEMENTS**.
- Assume that **THE INFORMATION IN THE SHORT STATEMENTS ARE TRUE**.

When making aspirin a chemist needed to make an ester bond. However, she only had access to a carboxylic acid. The carboxylic acid is less likely to form an ester bond than if an anhydride were used.

A) Reasonable Deduction B) Unreasonable Deduction

A) Reasonable Deduction B) Unreasonable Deduction

A) Reasonable Deduction B) Unreasonable Deduction

The most recent isolation of Taxol from the plant *Taxus brevifolia* had the highest percentage yield as compared to all the previous trials. The percentage yield measurements have a small uncertainty associated with them.

A) Reasonable Deduction B) Unreasonable Deduction

DRAWING CONCLUSION

A scientist will bring together several deductions, inferences or premises to arrive at a conclusion. The conclusion may form part of a larger argument. The ***strength of a conclusion is determined by how the deductions, inferences and/or premises support the conclusion.*** A **REASONABLE CONCLUSION** is ***related and relevant to the deductions, inferences or premises available*** and ***logically follows beyond a reasonable doubt*** (though it may not be certain). An **UNREASONABLE CONCLUSION** is poorly related or not relevant to the deductions, inferences or premises available.

Instructions:

- In this section you are presented with short statements followed by several hypotheses.
- You need to decide if each conclusion is a (A) Reasonable Conclusion or an (B) Unreasonable Conclusion.
- Base your choices **ONLY ON THE INFORMATION IN THE SHORT STATEMENTS.**
- Assume that **THE INFORMATION IN THE SHORT STATEMENTS ARE TRUE.**

The reactivity of a carbon centred molecule is related to the stability of its intermediate. The stability of the intermediate is correlated with the number of carbon atoms attached to its carbon centre. Adjacent carbon atoms can distribute the charge on the intermediate.

A) Reasonable Conclusion B) Unreasonable Conclusion

A) Reasonable Conclusion B) Unreasonable Conclusion

At rest a typical cell membrane has a concentration of 140 mM (milli molar) of potassium ions (K^+) within the cell. The concentration of potassium ions (K^+) within the cell can be decreased via open voltage gated potassium channels. The voltage gated potassium channels will only be open if the concentration of sodium ions (Na^+) is greater than 5 mM within the cell. At rest a typical cell membrane has concentration less than 5 mM of sodium ions (Na^+) within the cell.

A) Reasonable Conclusion B) Unreasonable Conclusion

A) Reasonable Conclusion B) Unreasonable Conclusion

A) Reasonable Conclusion B) Unreasonable Conclusion

ANALYSING ARGUMENTS

An argument may be the culmination of several conclusions. Scientist develop arguments based on their results or critique the arguments of others work. As a part of the scientific process scientists need to decide if an argument is valid or not. This requires identifying assumptions (stated or unstated), identifying inferences, deductions and premises and identifying conclusions.

A **VALID ARGUMENT** is based on *reasonable assumptions*. There is *adequate evidence* to support inferences, deductions and premises. This evidence comes from *credible sources*. The *conclusion are supported* the information presented. Finally the argument *is important and directly related* to the question being posed.

Instructions:

- In this section you are presented with short statements followed by several arguments.
- You need to decide if each argument is a **(A) Valid Argument** or an **(B) Invalid Argument**.
- Assume that THE INFORMATION IN THE SHORT STATEMENTS ARE TRUE.
- Assume that THE INFORMATION IN THE ARGUMENTS ARE TRUE.
- Avoid letting your personal biases influence your choices. Base your choices **ONLY ON THE INFORMATION IN THE SHORT STATEMENTS**.

The statement below is used for Questions 24 – 30.

Zinc oxide (ZnO) is the active ingredient in sunscreens which protects against DNA damage from UV radiation. Zinc oxide (ZnO) does this by effectively reflecting and diffracting harmful UV-radiation. Sunscreen residue left by tradespeople wearing sunscreen containing zinc oxide (ZnO₂) was thought to have caused rusting of many building materials. This has led lobbyists to be concerned that zinc oxide (ZnO₂) may in fact be toxic. Should it be removed from all sunscreens to decrease the risk of exposure to toxins?

24. Yes. Zinc oxide (ZnO) particles are smaller than 100nm (0.000000001 millimetres) and are small enough to be absorbed by skin cells. **While it is true that zinc oxide could be absorbed into cells there is no discussion to suggest that zinc oxide is a toxin.**

A) Valid Argument

B) Invalid Argument

25. Yes. There are other chemicals which could be used in sunscreens to absorb UV radiation before it can do damage to DNA. **While there are other chemicals that could absorb UV radiation, the argument is whether removing zinc oxide will reduce exposure to toxins, which is not discussed.**

A) Valid Argument

B) Invalid Argument

26. No. Removing zinc oxide (ZnO) would increase exposure to UV-radiation and therefore the risk of skin cancer. **This argument does not discuss the toxicity of zinc oxide, however reducing the damage to DNA caused by UV radiation is of significant importance making this a valid argument.**

A) Valid Argument

B) Invalid Argument

The statement below is used for Questions 27 – 30.

Bisphenol A (BPA) is a chemical used to manufacture re-useable polycarbonate water bottles. Bisphenol A (BPA) has been found to leech from water bottles into the water consumed. There is some controversy that bisphenol A (BPA) may be harmful towards humans. Should it be illegal to produce polycarbonate bottles containing bisphenol A (BPA)?

27. Yes. Lobby groups hypothesise that adverse health effects actually occur at doses far below levels previously established by standard toxicological procedures. **As this is a hypothesis there is no evidence to support this claim despite the argument being significant.**

A) Valid Argument

B) Invalid Argument

28. Yes. Bisphenol A (BPA) has a similar chemical structure to DNA. This could lead to Bisphenol A (BPA) interacting with DNA. **This could be considered a significant argument however there is no evidence to suggest that an interaction will occur or whether an interaction is detrimental.**

A) Valid Argument

B) Invalid Argument

29. No. Government and academic researchers estimate the intake of bisphenol A (BPA) is less than 0.00000125 milligrams per kilogram of body weight per day. The acceptable dose of bisphenol A (BPA) put forth by the US Environmental Protection Agency is 0.05 milligrams per kilogram of body weight per day. **This argument addresses a significant point and is supported by what would be considered credible sources.**

A) Valid Argument

B) Invalid Argument

30. No. Several independent studies from toxicology experts around the world were unable to replicate the finding of the research that showed that low doses of bisphenol A (BPA) led to adverse effects in mice. **This evidence suggests that the claims made by the University of Missouri are not reliable and there is no reason to remove BPA based on the Universities results.**

A) Valid Argument

B) Invalid Argument

Appendix I: DOT-CCTTv2 Statistics Summary

Table I1 Summary of descriptive statistics for DOT-CCTTv2 day 1 score, DOT-CCTTv2 day 2 score, and WTCGA-S score

		DOT-CCTTv2 day 1 Score	DOT-CCTTv2 day 2 Score	WGCTA-S Score
n		20	18	18
Mean		21.65	21.33	30.44
95% Confidence Interval for Mean	Lower Bound	20.74	19.54	27.54
	Upper Bound	22.56	23.12	33.35
Median		22.00	22.50	31.00
Variance		3.818	12.941	34.026
Standard Deviation		1.954	3.597	5.833
Minimum		18	13	18
Maximum		26	27	38
Range		8	14	20
Interquartile Range		2	5	11
Skewness		-.109	-.663	-.593
Kurtosis		.513	.112	-.622

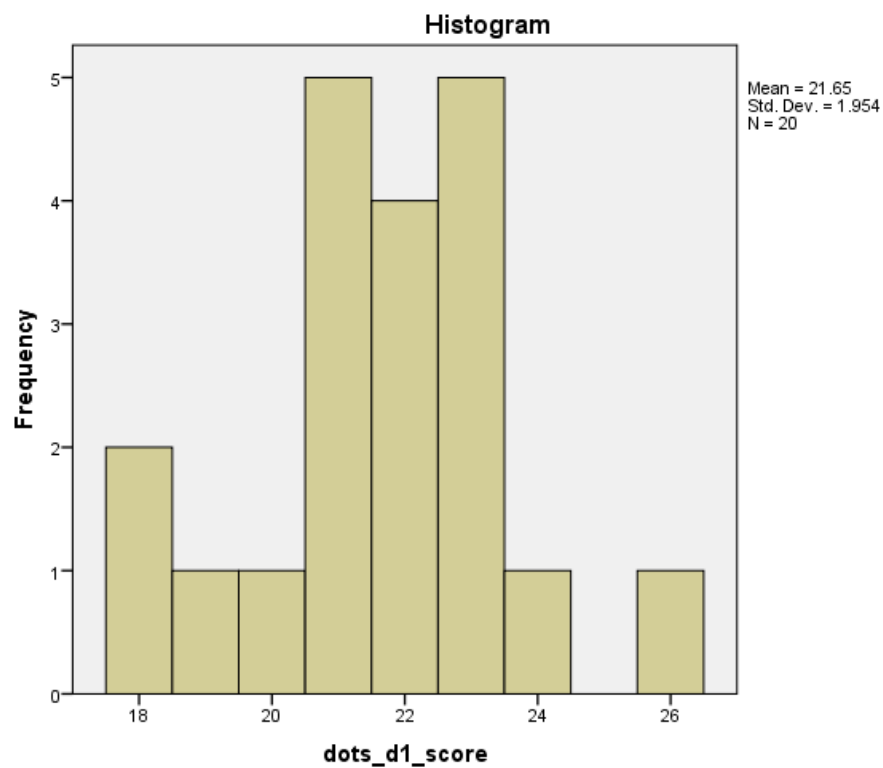


Figure I1 Histogram output from IBM SPSS of DOT-CCTTv2 day 1 scores (dots_d1_score) versus frequency

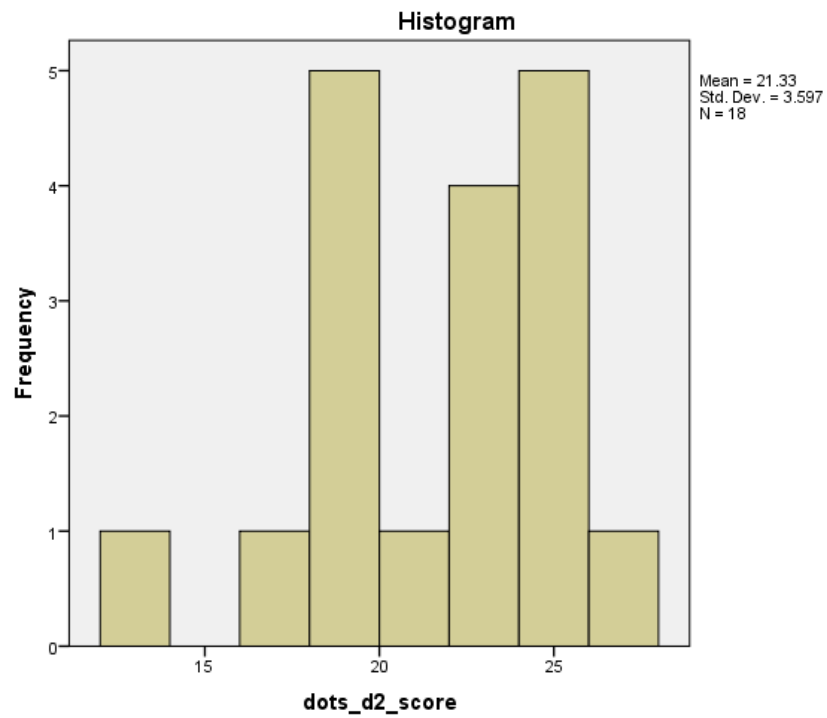


Figure I2 Histogram output from IBM SPSS of DOT-CCTTv2 day 2 scores (dots_d2_score) versus frequency

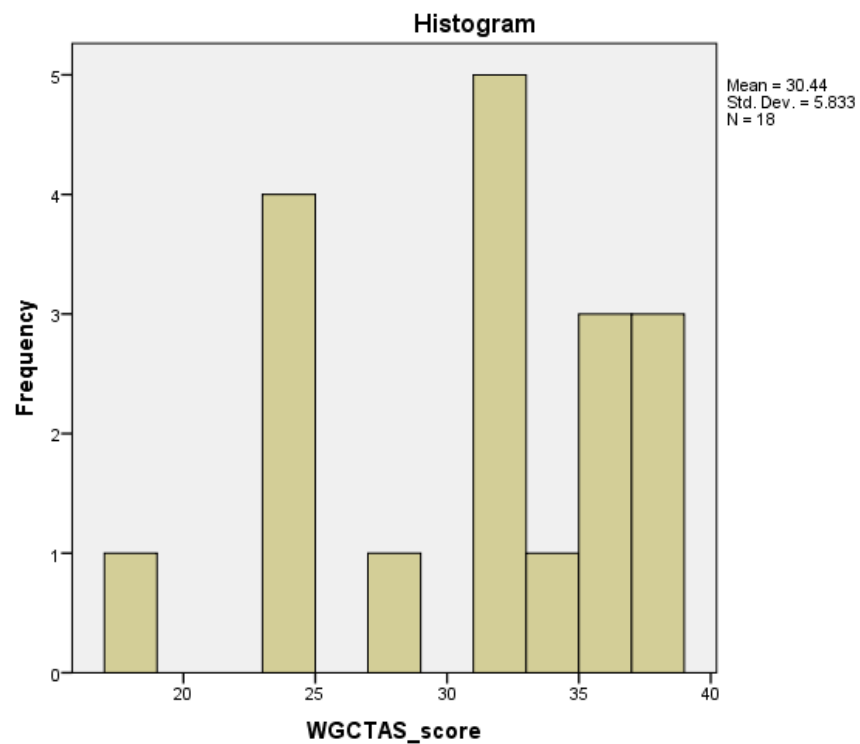


Figure I3 Histogram output from IBM SPSS of WGCTA-S scores (WGCTAS_score) versus frequency

Table 12 Summary of Wilcoxon Signed Rank Test comparing median score of the DOT-CCTTv2 on day 1 and day 2

Median DOT-CCTTv2 day 1 score	22.00
Median DOT-CCTTv2 day 2 score	22.50
Total number of cases	18
Z	-.114
Significance (<i>p</i>)	.909
Effect size (<i>r</i>)	.027

Table 13 Summary of Spearman's ρ correlation co-efficient comparing the median score of the DOT-CCTTv2 on day 1 versus the median score of the WGCTA-S

Median DOT-CCTTv2 day 1 score	22.00
Median WGCTA-S score	31.00
Total number of cases	18
Spearman's ρ	.314
Significance (<i>p</i>)	.205
Effect size (<i>r</i>)	.099

Table 14 Summary of Spearman's ρ correlation co-efficient comparing the median score of the DOT-CCTTv2 on day 1 and the median score of the WGCTA-S versus median ATAR

	DOT-CCTTv2 day 1 score	WGCTA-S score
Median ATAR	86.10	86.10
Total number cases	15	15
Spearman's ρ	.233	.466
Significance (<i>p</i>)	.403	.080
Effect size (<i>r</i>)	.054	.217

Appendix J: CCTTv2 Sample Student transcript and coding

Table J Codes and corresponding themes used in qualitative analysis of student transcripts

Codes	Themes
SCT	Strategies for completing the tests
DPK	Difficulties associated with prior knowledge
ACT	Awareness of bias and articulation of critical thought
EPL	Evidence of peer learning

Excerpt A

- S3: You would take the expectation from all your previous trials right? [EPL]
 S2: It wasn't expected. [EPL]
 S14: Because you don't actually know what's expected. [EPL]
 S3: That's true. [EPL]
 S14: They might have actually had higher expectation they might have actually used a new method- [EPL]
 S6: -new method for the synthesis that right- [EPL]
 S14: -but we might actually have a higher yield. [EPL]
 S10: What I was thinking like, let's say the previous yield was like 10 moles and the next one was like 10.001 moles it's technically still higher but is it really unexpected? [DPK]
 S6: But you think they would be – I would expect that they were constantly trying to improve your methods because you wanna make money or whatever and so you would kind of expect your results to – your yield to increase. [SCT]
 S11: But it never mentioned their expectations at all [SCT]
 S6: Yeah.
 S11: So they did it and they got a higher yield. [SCT]

Excerpt B

- S10: Just to note I think that this is a bit bias because like – like I think I might have drawn a bit from outside- [ACT/DPK]
 S14: -yeah I did, maybe actual learned that similar structures, similar function. [ACT/DPK]
 I: Okay, so what about the guys who put invalid argument for that one what do you say to that argument?
 S2: I don't know that information. [DPK]
 S14: Bisphenol is not the same structure as DNA though. [DPK]
 S10: Yeah, yeah, I know, I know. [DPK]
 I: Sorry?
 S14: I dunno I just thought that bisphenol A would not have the same structure as DNA. [DPK]
 S3: Kinda if you squint a lot might look DNA.
 (Laughter)
 S2: And maybe flip it upside down.
 S14: It says assume the information in the argument is actually true, so if actually I assume they actually have a similar structure they might be able have the same...
 S3: It reminds me of that thread that apparently said hey hang on the formula for sucrose is only off by a couple of carbons from meth therefore they're the same.
 S2: Where's the credible source from. That's what I looked at, or it's just a statement. If you read it off the internet doesn't mean it's true. [ACT]
 S14: But the second part of the statement is fine-
 S6: We're-we're told to assume that the information in the short statements is true- [SCT/ACT]
 S11: -it's not proven it just says that it could interact, it's not a proven fact. [SCT/ACT]

- S14: -it's the second part, the first part I just can't agree with it. [SCT/ACT]
 S6: -and the information in the arguments is true. It's like, I said because-its invalid cause it could lead to interacting but does that mean it should be illegal? [SCT/ACT]
 S2: Not even relating back I think that's in the examples as well. [SCT/ACT]

Excerpt C

- S8: Yeah it's talking about the level of attraction and the charge-
 S12: Yeah-
 S8: -of the complex
 S12: -to me, the fact that you know it starts talking about attraction rather than just the number of protons it can accept makes it fall outside of the information and is therefore an invalid assumption. [EPL/ACT/DPK]
 S8: Exactly, yeah. [EPL/ACT/DPK]
 I: So how many of you put valid assumptions? 1, 2...
 S13: I don't think it's right now. [EPL/ACT/DPK]
 I: Its fine, its fine. So why did you put valid assumptions in the first instance? [ACT/DPK]
 S13: I probably got caught up in the fact that most of the time, yes that's the case scientifically but we don't have that information that we can actually make that assumption. But I did any way with my own personal bias. [ACT/DPK]
 I: Okay. And seven of you, you put valid assumptions –ah invalid assumptions?
 S8: Yeah.
 S5: Yep.
 I: What's the reasoning for that?
 S7: Well it doesn't really mention the strength of attraction. It's only talking about what its formal charge is and how many things it can accept as a result. [SCT]
 S8: It's talking about the number of hydrogen ions, not the attraction towards the hydrogen ions. So there's no information regarding attraction we can't make an assumption. [SCT]
 S5: If you want to consider our actual chemistry knowledge – in fact we also – even if we do apply our prior chemistry knowledge to that it isn't quite – it still isn't a valid assumption because that negative charge could be distributed and have a lower charge density and therefore be less attractive than a more concentrated but lower total charge. [DPK]
 S12: Yeah, attraction and charge aren't directly proportional, it's just – you can't really predict one from the other so much. [DPK]

Excerpt D

- S4: I guess if you took out your bias and you were just reading it as a person who knew very little chemistry you would assume that failure means that it didn't form what you wanted it to form. So that's why I was like well that seems like a reasonable deduction to me. [SCT]
 S15: And I think also, sorry, it said that they were under the same reaction conditions. So you were kind of comparing like and like. Whereas I think question 14 you weren't really. It didn't specific it as much.
 I: And three of you put unreasonable deduction. So why is it?
 S16: Well I think it's because you don't know what the other conditions are, like what if they're really bad conditions for like carboxylic acid, it just says they're the same. They could be rush, like you don't know.
 S5: I said unreasonable deduction because it's a matter of failure. It said the carboxylic acid will probably, like on a one on one basis, the carboxylic acid will have like a lower probability form but overall the carboxylic acid can still form an ester group. So carboxylic acid plus the other stuff will still form aspirin so it – so I thought it was unreasonable to say the reaction would outright fail because you're using carboxylic acid instead of anhydride. It might not go as well it's not going to – I thought it was unreasonable to say it would outright fail. [ACT/DPK]

S12: Yeah and that the crux of what the whole question is based around, you know what your interpretation of failure is for this reaction. Is it not making any product, is it making less product- **[ACT/DPK]**

S5: Probability of less likely form.... **[ACT/DPK]**

S12: -is the anhydride less likely to form and that's-that one there is probably really based on interpretation, because of that. **[ACT/DPK]**

I: So we'll move to next section, drawing conclusion. So did you find any question in this section particularly difficult? Or annoying?

S12: 23 was interesting.

I: 23 right? Okay. Yes, so what is this asking?

S12: So we're looking at you know the concentration of sodium ions you know when the cell membrane is at rest. What really got me was I had to read it a second time 'cause it talks about gated sodium ions channels rather gates potassium ion channels, which is what the whole question is worded around. So, you know, to me that then after picking that up it falls outside of what we're looking at and becomes an unreasonable conclusion. **[EPL]**

I: so how many of you put unreasonable conclusion? 1, 2, 3 4, 5, 6, 7. And reasons you explained.

S5: It starts talking about something – **[EPL]**

S8: Yeah which the question doesn't say anything about sodium ions channels. Only the potassium ion channels and you can't assume they would have the same method of transporting across the cell membrane. **[EPL]**

S12: It states in the second last question there that voltage gated potassium channels open when the concentration of sodium ions increases but – yeah, no mention of the sodium ion channels. **[EPL]**

S5: Or even the existence of sodium gated ion channels. **[EPL]**

S12: Or even the existence of them. So it really falls outside the range of this question – ah outside the range of the information that we're given here. **[EPL]**

S4: Also they're only really talking about the information within the cell, the actual passage doesn't mention anything about outside the cell. **[EPL]**

S5: There's that too. **[EPL]**

I: And 2 of you said reasonable, whys that?

S13: I did not pick up on the fact that it is a sodium gated ion channel not potassium. **[EPL]**

I: Oh yeah?

S13: So I just assumed it was a potassium channel and from that simply thought, yeah that seems really reasonable.

S7: See the way I did it, which I think now probably wasn't the best way to read it was that it was gated by the sodium. So in my head I saw it as the same thing just with a different name.

Danczak-Overton-Thompson Chemistry Critical Thinking Test

Pilot (Version 1.3) 2017

By

Stephen Danczak, Chris Thompson and Tina Overton

Chemistry Education Research Group

Monash University, Australia



MONASH
Chemistry Education Research Group

This work is licensed under a Creative Commons
Attribution-NonCommercial-NoDerivatives 4.0 International License.



<https://creativecommons.org/licenses/by-nc-nd/4.0/>

DEMOGRAPHIC DATA

To enable the researchers to get the most useful data out of the test you are about to complete we'd greatly appreciate it if you completed the below section regarding demographic data.

With which sex do you identify?

☐ Female ☐ Male ☐ Other ☐ Rather not say

Which best describes your current level of study/employment

☐ 1st year undergraduate ☐ 2nd year undergraduate ☐ 3rd year undergraduate
☐ Honours ☐ Masters ☐ PhD
☐ Post-Doctoral ☐ Academic ☐ Other/Rather not say

What is your current age?

What is your preferred language?

What was your ATAR (or equivalent) score?

Thank you for your contribution.



MONASH
Chemistry Education Research Group

Introduction

Critical thinking is a term used to describe a set of cognitive abilities to arrive at desirable outcomes such as predicting and solving problems, innovation, developing and critiquing arguments. Critical thinking is a highly desirable trait sought after by employers around the world from a variety of disciplines (especially the sciences).

This test is designed to inform higher education practitioners and researchers as to how they can better design science programs aimed at developing your critical thinking based on the results of your cohort. Completion of this test will not affect your academic record in any way.

In the proceeding sections you will be asked 30 multiple choice questions, suitable for any level of chemistry knowledge, designed to assess five key areas of critical thinking:

MAKING ASSUMPTIONS
DEVELOPING HYPOTHESES
TESTING HYPOTHESES
DRAWING CONCLUSIONS
ANALYSING ARGUMENTS

Each section has its own set of instructions, so please read them carefully.

On behalf of your unit/subject/module coordinators and teaching staff we'd like to thank you for your contribution and hope you find the test interesting and challenging.

Regards,

The Chemistry Education Research Group (CERG)
Monash University, Australia

To the assessor,
Please note that the intended responses and the reasoning behind those responses are highlighted in green.

MAKING ASSUMPTIONS

An assumption is something that's taken for granted. Scientists generally clarify their assumptions in academic discussions. However some assumptions are implied. Some assumptions arise when the general consensus of the scientific community is that there is no additional proof or evidence required in the statement being made. Assumptions can be stated or implied.

For example:

A chemical company failed to fulfil a promise made in 2014 to develop innovations to reduce pollution because atmospheric carbon dioxide (CO₂) measured at the start of 2015 was above 350ppm.

It is a **VALID ASSUMPTION** that atmospheric carbon dioxide (CO₂) is an indicator of pollution.
It is an **INVALID ASSUMPTION** that the chemical company made no attempt to fulfil this promise.

Instructions:

- In this section you are presented with short passages followed by several assumptions.
- Select **(A) Valid Assumption** if you think the assumption **can be taken for granted** based on the information presented in the short passage.
- Select **(B) Invalid Assumption** if you think the assumption **cannot be taken for granted** based on the information presented in the short passage.
- *Treat each question individually* and base your decisions relating only to the original passage.
- For this section ask yourself would a person who has not studied science at university consider the assumption to be valid or invalid.

The passage below is used for Questions 1, 2, 3 and 4.

Metals which are paramagnetic or diamagnetic behave differently when exposed to an induced magnetic field. A chemist tested a metallic alloy sample containing thallium and lead by placing it in an induced magnetic field. From the test results the chemist decided the metallic alloy sample repelled the induced magnetic field and therefore was diamagnetic.

1. The thallium within the metallic alloy sample is responsible for determining how a metallic alloy interacts with an induced magnetic field. **There is no clear indication that the thallium is responsible for this interaction. It could be the lead or the combination of the two.**

A) Valid Assumption

B) Invalid Assumption

2. Paramagnetic metals do not repel induced magnetic fields. **The paragraph states that the alloy was deemed diamagnetic as it repelled the induced magnetic field. The paragraph states that that diamagnetic and paramagnetic metals behave differently therefore it is reasonable to assume that paramagnetic metals do not repel induced magnetic fields.**

A) Valid Assumption

B) Invalid Assumption

3. Metals can only ever be paramagnetic or diamagnetic. **There is no explicit mention of metals only being paramagnetic or diamagnetic and the paragraph does not make any attempt to suggest that these are the only two possibilities.**

A) Valid Assumption

B) Invalid Assumption

4. A diamagnetic metal alloy cannot be a paramagnetic metal alloy. **The paragraph suggests that diamagnetic metals and paramagnetic metals behave differently suggesting a metal cannot be both on the bases of the interactions with the induced magnetic field.**

A) Valid Assumption

B) Invalid Assumption

The passage below is used for Questions 5, 6 and 7.

Carbonate (CO_3^{2-}) has a formal charge of negative 2. Carbonate (CO_3^{2-}) can accept two hydrogen ions (H^+) which each have a formal charge of positive 1. Carbonate (CO_3^{2-}) is an example of a diprotic base. A monoprotic base such as hydroxide (OH^-) has a formal charge of negative 1. Hydroxide (OH^-) can only accept one hydrogen ion (H^+).

5. Carbonate (CO_3^{2-}) accepts hydrogen ions (H^+) more easily than monoprotic bases. **There is no mention of how readily either of the bases accept hydrogen ions.**

A) Valid Assumption

B) Invalid Assumption

6. Hydrogen ions (H^+) have greater attraction towards bases with a more negative formal charge. **There is no mention of a relationship between the degree on negative charge and hydrogen ion attraction.**

A) Valid Assumption

B) Invalid Assumption

7. The greater the formal negative charge of a base, the more hydrogen ions (H^+) it can accept. **This requires relating the two statements regarding how many hydrogen ion a base can accept. That is, recognising that base with a formal charge of negative 2 can accept 2 hydrogen ions (H^+) and that a base with a formal charge of negative 1 can only accept 1 hydrogen ion (H^+).**

A) Valid Assumption

B) Invalid Assumption

DEVELOPING HYPOTHESES

When generating hypotheses, scientists will draw inferences based on the data, observations and the supposed facts. An inference is used to fill in the gaps to create a connection, or look for the intended meaning. These inferences are not certain, but based on the information available, there is confidence in the hypothesis being developed.

For Example:

A chemist added one chemical to another and this was followed immediately by a colour change.

It is **LIKELY TO BE AN ACCURATE INFERENCE** that this occurred due to a reaction between the two chemicals. ***Based on the observations*** that the colour change occurred upon addition of the chemicals and the ***reasonable assumption*** that some reactions can result in a colour change this is **LIKELY TO BE AN ACCURATE INFERENCE**.

In the above example there would be **INSUFFICIENT INFORMATION TO DETERMINE THE ACCURACY** of a hypothesis which suggests one of the chemicals is more reactive than the other. ***More detailed observations are required***, for example reacting the chemicals in question with other chemicals.

If a chemist were to suggest that when two chemicals react they will always produce a colour change, this would be **LIKELY TO BE AN INACCURATE INFERENCE**. The ***conclusion is drawn from outside the information*** presented: they haven't tested other chemicals, and make an ***unreasonable assumption*** that all reactions result in a colour change.

Instructions:

- In this section you are presented with short passages followed by several inferences.
- You need to decide if each inference is/has:
 - A) Likely to be an accurate inference
 - B) Insufficient information to determine the accuracy
 - C) Likely to be an inaccurate inference
- *Treat each question individually* and base your decisions relating only to the original passage.

The passage below is used for Questions 8, 9 and 10.

The following is an excerpt adapted from an article posted by NASA at the URL: www.nasa.gov/press-release/nasa-confirms-evidence-that-liquid-water-flows-on-today-s-mars

Using an imaging spectrometer researchers detected signatures of hydrated minerals on slopes where mysterious streaks are seen on the red planet. These darkish streaks appear to ebb and flow over time. They darken and appear to flow down steep slopes during warm seasons, and then fade in cooler seasons. They appear in several locations on Mars when temperatures are above minus 23 degrees Celsius (minus 10 degrees Fahrenheit), and disappear at colder times.

“Our quest on Mars has been to ‘follow the water,’ in our search for life in the universe, and now we have convincing science that validates what we’ve long suspected,” said John Grunsfeld, astronaut and associate administrator of NASA’s Science Mission Directorate in Washington. “This is a significant development, as it appears to confirm that water -- albeit briny -- is flowing today on the surface of Mars.”

These downhill flows, known as recurring slope lineae (RSL), often have been described as possibly related to liquid water. The new finding may help to explain the dark features. The hydrated salts would lower the freezing point of a liquid brine, just as salt on roads here on Earth causes ice and snow to melt more rapidly. Scientists say it’s likely a shallow subsurface flow, with enough water wicking to the surface to explain the darkening.

8. It is a fact that water can exist in a liquid state below zero degrees Celsius due to the presence of the hydrated salts. On Mars this could lead to the freezing and thawing of water resulting in the ebb and flow of the mysterious streaks on Mars. **This analogy is discussed in the third paragraph and it would be reasonable to assume that this natural phenomena can occur on Earth and Mars.**

- A) Likely to be an accurate inference
- B) Insufficient information to determine the accuracy
- C) Likely to be an inaccurate inference

9. Some planets are likely to have similar conditions to Mars and NASA is planning to determine if there may be habitable water on these other planets. **While the second paragraph mentions ‘in our search for life in the universe’ and it is reasonable to assume a group like NASA are actively pursuing this goal, the assumption is drawn from outside what is presented and more information would be required to decide if there were plans in place.**

- A) Likely to be an accurate inference
- B) Insufficient information to determine the accuracy
- C) Likely to be an inaccurate inference

10. The imaging spectrometer directly detects water on Mars. **There was no mention that water was measured directly. This conclusion is drawn from outside what is presented. The first paragraph describes that hydrated minerals were detected via spectrometry not water itself.**

- A) Likely to be an accurate inference
- B) Insufficient information to determine the accuracy
- C) Likely to be an inaccurate inference

The passage below is used for Questions 11, 12 and 13.

The following is an excerpt adapted from an article posted by CSIRO at the URL:

www.csiro.au/en/News/News-releases/2015/Marine-debris

Researchers from CSIRO and Imperial College London have assessed how widespread the threat of plastic is for the world's seabirds and found the majority of species have plastic in their gut.

The study, led by Dr Chris Wilcox with co-authors Dr Denise Hardesty and Dr Erik van Seville and published today in the journal PNAS, found that nearly 60 percent of all seabird species have plastic in their gut. Based on analysis of published studies since the early 1960s, the researchers found that plastic is increasingly common in seabirds' stomachs.

In 1960, plastic was found in the stomach of less than 5 per cent of individual seabirds, rising to 80 per cent by 2010. The researchers predict that plastic ingestion will affect 99 per cent of the world's seabird species by 2050, based on current trends. The scientists estimate that 90 per cent of all seabirds alive today have eaten plastic of some kind.

This includes bags, bottle caps, and plastic fibres from synthetic clothes, which have washed out into the ocean from urban rivers, sewers and waste deposits. Birds mistake the brightly coloured items for food, or swallow them by accident, and this causes gut impaction, weight loss and sometimes even death.

11. This passage implies that as a result of 90 percent of all seabirds having eaten plastic of some kind, some species of sea bird will become extinct. **The fourth paragraph states sometimes death occurs in birds which have consumed plastic however to suggest this would lead to extinction is a conclusion drawn from outside the passage.**

- A) Likely to be an accurate inference
- B) Insufficient information to determine the accuracy
- C) Likely to be an inaccurate inference**

12. The number of seabirds to have eaten plastic of some kind has a strong link to pollution caused by consumer and industrial waste. **While it not certain there is explicit mention of plastics such as bags, bottle caps and clothing washing out to sea and it is reasonable to assume these are forms of consumer and industrial waste.**

- A) Likely to be an accurate inference**
- B) Insufficient information to determine the accuracy
- C) Likely to be an inaccurate inference

13. The increase in the number of coloured plastic items is responsible for a decline in the number of seabirds which haven't eaten plastic. **There is discussion in the final paragraph of birds mistaking brightly coloured plastics for food however to assume a causation between the two would require additional information.**

- A) Likely to be an accurate inference
- B) Insufficient information to determine the accuracy**
- C) Likely to be an inaccurate inference

TESTING HYPOTHESES

Scientists conduct experiments to test hypotheses. They begin with a hypothesis or statement they believe to be true and systematically seek information to confirm or refute the hypothesis. This results in a premise which is believed to be accurate or true.

For Example:

After completing a synthesis, a chemist compared the compound produced to a specific property of the starting material. The hypothesis is that if the specific property of the compound produced is different to that of the starting material there is no starting material in the compound produced. The chemist finds that the compound produced is very different with respect to this specific property of the starting material.

A **REASONABLE DEDUCTION** is that there is no starting material in the product. The *evidence presented is supported by this hypothesis*.

An **UNREASONABLE DEDUCTION** would be that based on these observation the compound produced is certainly the chemical the chemist was intending to produce. *The deduction is not related to the hypothesis* and there is *insufficient evidence to support this claim*.

Instructions:

- In this section you are presented with short passages followed by several hypotheses.
- You need to decide if each hypothesis is a **(A) reasonable deduction** or an **(B) unreasonable deduction**.
- Base your choice only on the information in the short passages.
- Assume that the information in the short passages are true.
- *Treat each question individually* and base your decisions relating only to the original passage.

The passage below is used for Questions 14, 15 and 16.

A chemist needed to make aspirin using a carboxylic acid. One of the by-products of this chemical reaction is water. The carboxylic acid is less likely to form aspirin when the water by-product is not removed during the chemical reaction.

14. Using a carboxylic acid without removing water during the chemical reaction will produce an insignificant amount of aspirin. **The passage refers to likelihood of formation of aspirin and this cannot necessarily be related to the amount of aspirin produced.**

A) Reasonable Deduction

B) Unreasonable Deduction

15. Removing water during the chemical reaction is more likely to succeed in forming aspirin than when water is not removed during the chemical reaction. **The passage clearly states that the carboxylic acid will be less successful likely to form aspirin when water is not removed.**

A) Reasonable Deduction

B) Unreasonable Deduction

16. Assuming all other reaction conditions are the same, the formation of aspirin using a carboxylic acid without removing water has a higher probability of aspirin not forming than if water were removed. **The passage clearly states that the carboxylic acid will be less successful likely to form aspirin when water is not removed. It is essentially a rewording of the previous question asking what the relationship between the removal of water and the formation of aspirin is.**

A) Reasonable Deduction

B) Unreasonable Deduction

The passage below is used for Questions 17 and 18.

The most recent extraction of the chemical Taxol from the plant *Taxus brevifolia* resulted in the highest amount of Taxol extracted from the plant compared to all the previous trials. The amounts of Taxol extracted have some degree of uncertainty associated with the measurements.

17. The amount of Taxol extracted in the most recent extraction was higher than expected. **There is no mention of the expectation.**

A) Reasonable Deduction

B) Unreasonable Deduction

18. The amount of Taxol extracted in the most recent extraction was greater than first extraction. **The passage states that the amount of Taxol extracted was the highest of the previous six trials.**

A) Reasonable Deduction

B) Unreasonable Deduction

DRAWING CONCLUSIONS

A scientist will bring together several pieces of information including deductions, inferences or premises to arrive at a conclusion. The conclusion may form part of a larger argument. The ***strength of a conclusion is determined by how the deductions, inferences and/or premises support the conclusion.***

For example:

In a series of ten identical iron sheets, five of the sheets were galvanised with metallic zinc. All ten sheets were then exposed to high levels of oxygen for 24 hours. The sheets which were not galvanised with metallic zinc quickly rusted whereas the sheets galvanised with metallic zinc did not.

A **REASONABLE CONCLUSION** is that exposure to oxygen likely caused the sheets which were not galvanised with metallic zinc to quickly rust. The deduction is **based on relevant information** and the logic **follows beyond a reasonable doubt** (though it is not certain).

An **UNREASONABLE CONCLUSION** is that galvanising with other metals will also prevent iron panels from rusting. The conclusion is **not directly supported by relevant information** and **does not follow beyond a reasonable doubt**.

Instructions:

- In this section you are presented with short passages followed by several conclusion.
- You need to decide if each conclusion is a (A) Reasonable Conclusion or an (B) Unreasonable Conclusion.
- Base your choices only on the information in the short passages.
- Assume that the information in the short passages are true.
- *Treat each question individually* and base your decisions relating only to the original passage.

The passage below is used for Questions 19 and 20.

When a chemical reaction can occur between two or more molecules the reactivity of those molecules is related to the stability of the transition state of the molecules. The stability of transition states is correlated with activation energy. Catalysts can lower the activation energy of transition states.

19. If the transition state is very stable the molecules are very reactive. **There is no mention of a positive correlation between reactivity and stability, just that there is a correlation.**

A) Reasonable Conclusion

B) Unreasonable Conclusion

20. If the activation energy of a transition state is lowered it is always due to a catalyst. **There is insufficient information to suggest catalyst are the only way to lower the activation energy of a transition state.**

A) Reasonable Conclusion

B) Unreasonable Conclusion

The passage below is used for Questions 21, 22 and 23.

At rest a typical cell has a concentration of 140 mM (milli molar) of potassium ions (K^+) within the cell. The concentration of potassium ions (K^+) within the cell can be decreased via open potassium channels. The potassium channels will only be open if the concentration of sodium ions (Na^+) is greater than 5 mM within the cell. At rest a typical cell has a concentration of less than 5 mM of sodium ions (Na^+) within the cell.

21. Assume that the potassium channels of a typical cell at rest are not open. In this situation, when the concentration of sodium ions (Na^+) increases outside of a typical cell at rest, sodium ion channels will open. **There is no information to suggest that sodium ion channels open or close due to a change in concentration of sodium ions outside the cell.**

A) Reasonable Conclusion

B) Unreasonable Conclusion

22. At rest the potassium channels of a typical cell are closed. **The passage clearly states that potassium channels are closed if the concentration of sodium is greater than 5 mM. The passage then goes on to state that the concentration of sodium in a cell at rest is 5 mM. Therefore at rest the potassium ion gate is closed.**

A) Reasonable Conclusion

B) Unreasonable Conclusion

23. At rest the number of potassium channels open on a cell cannot be determined. **As stated above at rest all the potassium channels are closed at rest. Therefore the number of channels open is zero and can be determined.**

A) Reasonable Conclusion

B) Unreasonable Conclusion

ANALYSING ARGUMENTS

As a part of the scientific process scientists need to decide if an argument is valid or not. This requires identifying assumptions (stated or unstated), identifying inferences, deductions and premises, identifying conclusions (some conclusions in a statement may be implied) and if the argument is related to the question being posed. An argument can be considered weak even if there is adequate evidence, credible sources and supporting information is provided BUT the argument is not important AND directly related to the question being posed).

For example:

There are concerns that harmful pesticides may be present in the food supply due to commercial agriculture. Should the Australian government introduce regulations to decide which industrial pesticides are safe for commercial agriculture to prevent the consumption of harmful chemicals?

Yes, the European Union has prohibited the use of possible cancer causing chemicals such as Atrazine, however these chemicals have been found in effluent water of Australian agriculture.

This is a **VALID ARGUMENT** as it is **reasonable to assume** the harmful pesticides may be consumed, the European Union are a **reliable source of information** and cancer causing chemicals are **important and directly related to the question**.

No, as genetically modified organisms (GMOs) become more prevalent the dependence on pesticides naturally declines. For example GMO cotton expressing bacterial toxins to target specific pests have reduced the use of specific pesticides by 87% since the 90s.

This is an **INVALID ARGUMENT**. While it is **reasonable to assume** GMOs reduce the dependence on pesticides **it is not relevant** as it does not address the fact the some pesticides are harmful. Also it is **not reasonable to assume** that the reductions in pesticide use in cotton manufacture can be generalised to other crops.

Instructions:

- In this section you are presented with short passages followed by several arguments.
- You need to decide if each argument is a **(A) Valid Argument** or an **(B) Invalid Argument**.
- Assume that the information in the short passages are true.
- Assume that the information in the arguments are true.
- Avoid letting your personal biases influence your choices. Base your choices **only on the information in the short statements**.
- *Treat each question individually* and base your decisions relating only to the original passage.

The passage below is used for Questions 24, 25 and 26.

Zinc oxide (ZnO) is the active ingredient in sunscreens, which protects against DNA damage from UV radiation. Zinc oxide (ZnO) does this by effectively reflecting and diffracting harmful UV-radiation. Sunscreen residue left by tradespeople wearing sunscreen containing zinc oxide (ZnO) was thought to have caused rusting of many building materials. This has led many people to be concerned that zinc oxide (ZnO) may be toxic. Should it be removed from all sunscreens to decrease the risk of exposure to toxins?

24. Yes. Zinc oxide (ZnO) particles in sun screens are smaller than 100 nm (0.000000001 millimetres) and behave differently at such a small scale. **While it is true that zinc oxide behaves differently, there is no discussion to suggest that zinc oxide is a toxin and this point is unrelated to the argument.**

A) Valid Argument

B) Invalid Argument

25. Yes. There are other chemicals which absorb UV radiation. **While there are other chemicals that could absorb UV radiation, the argument is whether removing zinc oxide will reduce exposure to toxins, which is not discussed here.**

A) Valid Argument

B) Invalid Argument

26. No. Removing zinc oxide (ZnO) would increase exposure to UV-radiation and therefore the risk of skin cancer. **This argument does not discuss the toxicity of zinc oxide, however reducing the damage to DNA caused by UV radiation is of significant importance making this a valid argument.**

A) Valid Argument

B) Invalid Argument

The passage below is used for Questions 27, 28, 29 and 30.

Bisphenol A (BPA) is a chemical used to manufacture re-useable polycarbonate water bottles. Bisphenol A (BPA) has been found to leech from water bottles into the water consumed. There is some controversy that bisphenol A (BPA) may be harmful towards humans. Should it be illegal to produce polycarbonate bottles containing bisphenol A (BPA)?

27. Yes. Many parent groups hypothesise that adverse health effects actually occur at doses far below levels established by toxicology experts. **As this is a hypothesis there is no evidence to support this claim despite the argument being significant. It's true that these groups hypothesise this but that does not make the hypothesis true.**

A) Valid Argument

B) Invalid Argument

28. Yes. Bisphenol A (BPA) is sometimes used in baby bottles. This could lead to Bisphenol A (BPA) being consumed by babies. **This could be considered a significant argument however there is no evidence to suggest that an interaction will occur or whether an interaction is detrimental.**

A) Valid Argument

B) Invalid Argument

29. No. Government and academic researchers estimate the intake of bisphenol A (BPA) is less than 0.00000125 milligrams per kilogram of body weight per day. The acceptable dose of bisphenol A (BPA) put forth by the US Environmental Protection Agency is 0.05 milligrams per kilogram of body weight per day. **This argument addresses a significant point and is supported by what would be considered credible sources.**

A) Valid Argument

B) Invalid Argument

30. No. Over fourteen unbiased studies carried out by toxicology experts from research universities around the world were unable to replicate the finding of the research paper that showed that low doses of bisphenol A (BPA) led to adverse effects in mice. **The argument suggests that essentially the original study has attempted to be replicated by several credible sources and the results were not reproducible.**

A) Valid Argument

B) Invalid Argument

Appendix L: DOT-CCTTv3 Statistics Summary

Table L1 Summary of descriptive statistics for DOT-CCTTv3 score

		DOT-CCTTv3 Score
n		298
Mean		19.16
95% Confidence Interval for Mean	Lower Bound	18.58
	Upper Bound	19.75
Median		20.00
Variance		26.057
Standard Deviation		5.105
Minimum		5
Maximum		30
Range		25
Interquartile Range		8
Skewness		-.168
Kurtosis		-.719

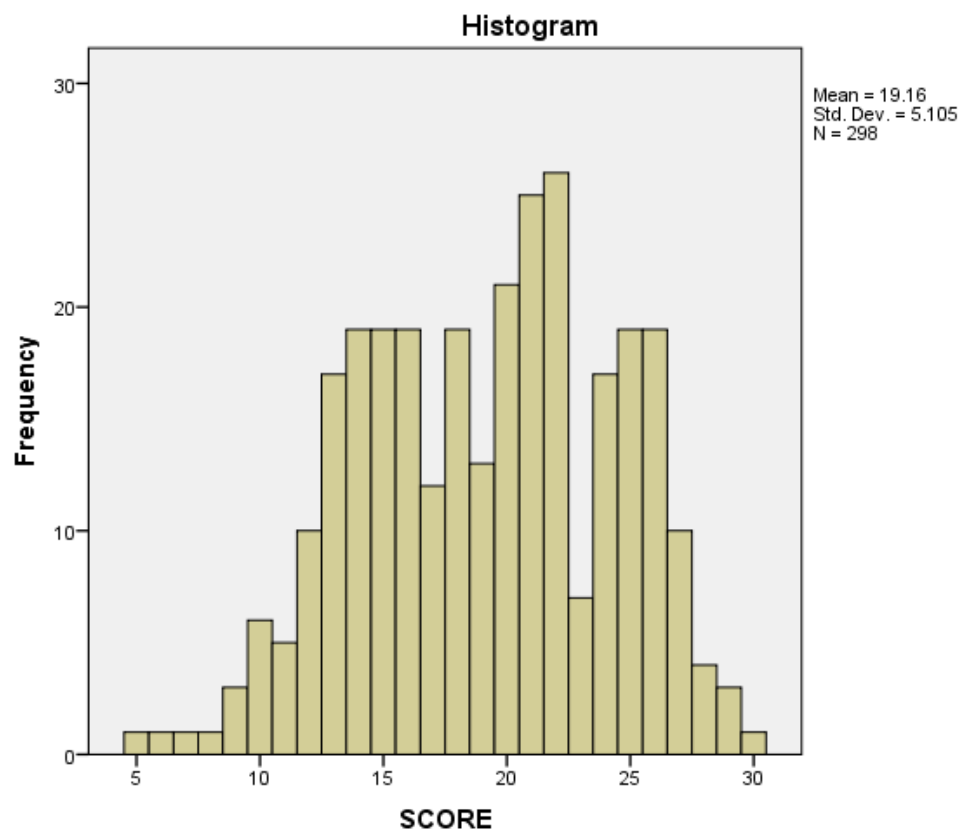


Figure L Histogram output from IBM SPSS of DOT-CCTTv3 scores versus frequency

Table L2 Scale reliability of DOT-CCTTv3 score treating the DOT-CCTTv3 as a single scale of 30 items

Cronbach's α	Number of Items including scale	Cases	Number of Cases	Percentage of Cases (%)
.710	31	Valid	266	89.3
		Excluded ^a	32	10.7
		Total	298	100.0

^aListwise (cases with data missing from the items of interest) deletion based on all variables in the procedure.

Table L3 Scale reliability of DOT-CCTTv3 score treating the DOT-CCTTv3 as single scale made of five sub-scales

Cronbach's α	Number of Items including scale	Cases	Number of Cases	Percentage of Cases (%)
.780	6	Valid	266	89.3
		Excluded ^a	32	10.7
		Total	298	100.0

^aListwise (cases with data missing from the items of interest) deletion based on all variables in the procedure

Table L4 Summary of correct item total correlations and Cronbach's α if item deleted of the five sub-scales of the DOT-CCTTv3 score

Sub-scale	Corrected Item-Total Correlation	Cronbach's α if Item Deleted
Making Assumptions	.709	.729
Developing Hypotheses	.522	.764
Testing Hypotheses	.577	.760
Drawing Conclusions	.602	.757
Analysing Arguments	.717	.720

Table L5 Example reliability of sub-scale score 'Making Assumptions'

Cronbach's α	Number of Items including sub-scale	Cases	Number of Cases	Percentage of Cases (%)
.682	8	Valid	297	99.7
		Excluded ^a	1	.3
		Total	298	100.0

^aListwise (cases with data missing from the items of interest) deletion based on all variables in the procedure

Table L6 Summary of correct item total correlations and Cronbach's α if item deleted of questions in the sub-scale 'Making Assumptions'

Questions	Corrected Item-Total Correlation	Cronbach's α if Item Deleted
Q1	.491	.639
Q2	.366	.659
Q3	.140	.693
Q4	.292	.670
Q5	.492	.638
Q6	.403	.651
Q7	.352	.668

Table L7 Summary of split halves reliability for the DOT-CCTTv3

Cronbach's α	Part 1	α	.625
		Number of Items	15 ^a
	Part 2	α	.624
		Number of Items	15 ^b
	Total Number of Items		30
Correlation Between Part 1 and Part 2			.666
Spearman-Brown Coefficient (ρ_{cc})			.800

^a The items are: Q1, Q3, Q5, Q7, Q9, Q11, Q13, Q15, Q17, Q19, Q21, Q23, Q25, Q27, Q29.

^b The items are: Q2, Q4, Q6, Q8, Q10, Q12, Q14, Q16, Q18, Q20, Q22, Q24, Q26, Q28, Q30.

Table L8 Sample Mann-Whitney U comparing median DOT-CCTTv3 scores of participants who answered Question 1 of the sub-scale 'Making Assumption' correctly and those who answered it incorrectly

Question 1	Correct	Number of Cases	200
		Median	21
	Incorrect	Number of Cases	97
		Median	15
	Total number of Cases		297
Mann-Whitney U			16129.00
Standardised Test Statistic (z score)			9.280
Significance (p)			.000
Effect size (r)			.54

Table L9 Sample Mann-Whitney U comparing median DOT-CCTTv3 scores of the first year education group *versus* median DOT-CCTTv3 score of the postgraduate education group

Education Group	First Year	Number of Cases	119
		Median	16
	Postgraduate	Number of Cases	80
		Median	24
	Total number of Cases		199
Mann-Whitney U			8338.00
Standardised Test Statistic (z score)			9.000
Significance (p)			.000
Effect size (r)			.64

Table L10 Summary of Spearman's ρ correlation co-efficient comparing the median DOT-CCTTv3 scores of third year Monash University students *versus* the median DOT-CCTTv3 scores of third year Curtin University students

University Attended	Monash University	Number of Cases	44
		Median	20
	Curtin University	Number of Cases	24
		Median	22
	Total number of Cases		68
Mann-Whitney U			675.00
Standardised Test Statistic (z score)			1.835
Significance (p)			.067
Effect size (r)			.22

Table L11 Summary of Spearman's ρ correlation co-efficient comparing the median score of the DOT-CCTTv3 *versus* median ATAR and median age

		Age	ATAR
DOT-CCTTv3 Score	Median	20	90.20
	Total number cases	284	194
	Spearman's ρ	.504	.197
	Significance (p)	.000	.006
	Effect size (r)	.254	.039