

The Impact of Test Items Incorporating Multimedia Stimuli on the Performance and Attentional Behaviour of Test-Takers

Paula Lehane

B.Ed Psych (Hons), H.Dip (SEN), M.Ed (ASN)

Dissertation submitted to Dublin City University in fulfilment of the requirements for the award of Doctor of Philosophy.

Supervisors: Prof Michael O'Leary, Dr Darina Scully and Prof Mark Brown

Dublin City University
School of Policy and Practice

December 2021

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, and that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: *Paula Kehoe*

ID No.: 18211410

Date: 17th December 2021

Table of Contents

Table of Contents	ii
List of Tables.....	ix
List of Figures	xii
List of Acronyms and Abbreviations	xvi
Acknowledgements	xvii
Abstract	xix

Chapter 1: Introduction

1.1 Research Topic and Problem	1
1.2 Significance of the Study	2
1.2.1 Building 'good' items: Understanding their features and characteristics.....	3
1.2.2 Selecting the right stimulus: Images, Animations and Simulations	5
1.2.3 Leveraging process data to inform item design and validity claims.....	7
1.3 Origins and Rationale for the Study.....	8
1.4 Scope of the Study	9
1.5 Organisation of the Thesis.....	11

Chapter 2: Literature Review

2.1 Introduction.....	13
2.2 Scope of the Literature Review	13
2.3 Technology Based Assessments (TBAs)	14
2.4 Technology-Based Items.....	15
2.4.1 Classifying Technology-Based Items	15
2.4.2 Selected Response (SR) Items	19
2.4.3 Figural Response (FR) Items.....	23
2.4.4 Constructed Response (CR) Items	28

2.4.5 Simulation-Type Items and Tasks	29
2.5 Commentary on Technology Based Items	31
2.5.1 TEI Utility Framework (Russell, 2016)	34
2.6 Towards a Cognitive Theory of Multimedia Assessment.....	36
2.6.1 Cognitive Theory of Multimedia Learning (CTML)	37
2.6.2 Applying the CTML to assessment.....	41
2.6.2.1 The Role of Cognitive Load	41
2.6.2.2 Construct Measurement	42
2.6.2.3 Expertise Reversal Effect.....	44
2.6.3 A Cognitive Theory of Multimedia Assessment (CTMA)	46
2.6.3.1 Evaluating the CTMA	50
2.7 Advancing the field of TBAs.....	51
2.7.1 Value of Response Process Data	53
2.7.1.1 Eye-tracking Technology and Eye Movement Data	55
2.8 Summary and Conclusions.....	58

Chapter 3: Methodology

3.1 Introduction.....	61
3.2 Conceptual Framework.....	61
3.3 Research Questions.....	65
3.4 Research Design	68
3.4.1 Study 1 and Study 3: Mixed Methods Factorial Design	68
3.4.2 Study 2 and Study 3: Exploratory Eye-Tracking Study	71
3.5 Research Participants and Sampling.....	72
3.6 Instrumentation	73
3.6.1 TBA of Scientific Literacy – Study 1 (A, B)	73
3.6.1.1 Classifying Test Items for a TBA of Scientific Literacy (PISA 2015) ...	75

3.6.1.2 Creating a TBA for Scientific Literacy: Static and Dynamic Stimuli	77
3.6.1.3 Testing Platform	79
3.6.2 <i>Simulation-Type Items – Study 2</i>	81
3.6.2.1 Classifying Simulation-Type Items	82
3.6.3 <i>Cognitive Interview – Study 3</i>	83
3.7 Equipment	84
3.8 Measures and Variables	85
3.8.1 <i>Performance Data</i>	86
3.8.1.1 Study 1 (A, B) – Scores on a TBA of Scientific Literacy	86
3.8.1.2 Study 2 – Scoring Simulation-Type Items	86
3.8.2 <i>Process Data</i>	87
3.8.2.1 Times of Interest (TOIs)	88
3.8.2.2 Areas of Interest (AOIs)	90
3.8.2.3 Eye Movement Metrics	95
3.8.3 <i>Qualitative Data</i>	98
3.8.4 <i>Demographics</i>	98
3.9 Pilot Study	99
3.9.1 <i>Pilot Study 1: Testing Platform</i>	99
3.9.2 <i>Pilot Study 2: Eye-Tracking</i>	100
3.10 Main Study	102
3.10.1 <i>Study 1A</i>	102
3.10.2 <i>Study 1B, Study 2 and Study 3</i>	102
3.11 Ethical Considerations	103
3.12 Data Analysis	104
3.13 Summary	106

Chapter 4: Results

4.1 Introduction	107
4.2 Study 1A	108
4.2.1 Demographics	108
4.2.2 Research Question 1 (RQ1): Do different multimedia stimuli (e.g. images, animations) affect test-taker performance in a TBA of scientific literacy?.....	110
4.2.2.1 RQ1a: Is the performance of test-takers on items in a TBA of scientific literacy affected by the type of multimedia stimulus used?.....	111
4.2.2.2 RQ1b: Is the performance of test-takers in a TBA affected by the type of multimedia stimulus used when their previous levels of prior knowledge are considered?	112
4.2.2.3 RQ1c: Does the type of multimedia stimulus used affect key item statistics?	113
4.2.3 Summary: Study 1A	117
4.3 Study 1B	118
4.3.1 Demographics	118
4.3.2 Research Question 2 (RQ2): How do different multimedia stimuli (e.g. images, animations) affect the attentional behaviour of test-takers in TBAs?.....	119
4.3.2.1 RQ2a: Does the number of visits to an item's interaction space differ according to the multimedia object used?	119
4.3.2.2 RQ2b: Does the average duration of whole fixations in the interaction space of an item differ according to the multimedia stimulus used?	120
4.3.2.3 RQ2c: Does the proportion of fixations to the interaction space of an item differ according to the multimedia object used?	122
4.3.3 Summary: Study 1B	122
4.4 Study 2	123
4.4.1 Demographics and Performance	123
4.4.2 Research Question 3 (RQ3): What behaviours are demonstrated by test-takers when responding to items and tasks involving simulations?.....	125
4.4.2.1 RQ3a: Is time-on-task and time-per-phase related to test-taker performance?.....	125
4.4.2.2 RQ3b: What relationship, if any, does the number of simulations run per task have on test-taker performance?	128

4.4.2.3 RQ3c: What attentional behaviours (<i>number of visits</i>) do test-takers exhibit when completing simulation type items in the Orientation phase?	129
4.4.2.4 RQ3d: What attentional behaviours (<i>time-to-first-fixation, number of whole fixations, proportion of fixations</i>) do test-takers exhibit when completing simulation type items in the Output phase?	131
4.4.3 Summary: Study 2.....	138
4.5 Study 3	139
4.5.1 Demographics.....	139
4.5.2 Familiarisation	143
4.5.3 Sense-Making.....	144
4.5.3.1 Information Gathering	145
4.5.3.2 Identifying Relevant Information	146
4.5.4 Making Decisions	148
4.5.4.1 Pre-Decision Strategies	149
4.5.4.2 Post-Decision Checks	150
4.5.5 Feedback.....	151
4.5.6 Summary: Study 3.....	153
4.6 Summary	154

Chapter 5: Discussion and Conclusions

5.1 Introduction.....	155
5.2 Summary of Research	155
5.3 Key Findings	156
5.3.1 Static and Dynamic Objects as Item Stimuli.....	156
5.3.1.1 Impact on test-taker performance and item functioning	156
5.3.1.2 Impact on test-taker behaviour	158
5.3.2 Engaging with Simulation-Type Items.....	159
5.3.3 Interacting with Technology-Based Items	162

5.3.3.1 Interviewee 8	164
5.3.3.2 Test-Takers' views on TBAs in post-primary settings	165
5.4 Conclusions	165
5.5 Study Strengths and Limitations	168
5.6 Theoretical and Practical Implications	171
5.7 Recommendations.....	173
5.7.1 Policy and Practice.....	173
5.7.2 Future Research	175
5.8 Epilogue.....	177
References	179
 Appendix A: Assumptions of the Cognitive Theory of Multimedia Learning	 198
A1. Dual Channel Assumption	198
A2. Limited Capacity Assumption	199
A3. Active Processing Assumption	200
Appendix B: Programme for International Student Assessment (PISA)	201
B1. Programme for International Student Assessment (PISA): An Overview	201
B2. Understanding PISA's Scientific Literacy Framework	202
B3. Scientific Literacy: Test Items	204
Appendix C: Items in Static and Dynamic Conditions (Study 1A, Study 1B).....	206
Appendix D: Introductory Screens of Testing Platform	219
Appendix E: Study 2 Materials	220
Appendix F: Designing an Eye-Tracking Study	224
F1. Ensuring Data Quality	224
Appendix G: Interview Schedule: cued-Retrospective Think-Aloud (cRTA)	227
Appendix H: Survey	228

Appendix I: Standardised Instructions for Study 1A, Study 1B	229
Appendix J: DCU Research Ethics Committee Approval	232
Appendix K: Study 3 Analysis: Step 2 (Initial Coding)	233
Appendix L: Study 3 Analysis: Step 3 (Theme Search)	234
Appendix M: Study 3 Analysis: Step 4/Step 5 (Theme Review/Definition)	235

List of Tables

Table 2.1 Three assumptions of the Cognitive Theory of Multimedia Learning (CTML)	39
Table 2.2 Ten design principles for multimedia methods of instruction in computerised learning environments (Mayer, 2017)	40
Table 2.3 Forms of cognitive assessment load	46
Table 2.4 Adapting the CTML for Assessment (adapted from Kirschner et al., 2016, p. 30)	49
Table 3.1 Controls for threats of internal validity (adapted from Howitt & Cramer, 2008; Creswell, 2014; Coleman, 2019).....	71
Table 3.2 Units used in Study 1	77
Table 3.3 Fixation and visit metrics for TOIs and AOIs	96
Table 3.4 Descriptors for reporting and interpreting effect sizes	105
Table 3.5 Braun and Clark's (2006) six-step framework for thematic analysis	105
Table 4.1 Demographic Details: Study 1A	108
Table 4.2 Percentages of students agreeing/disagreeing with statements about their enjoyment of learning science.....	109
Table 4.3 Performance of participants involved in Study 1A	111
Table 4.4 Frequencies of correct answers on individual items: Static, Dynamic	112
Table 4.5 Scores of participants in Study 1A according to condition and levels of prior science knowledge	113
Table 4.6 Values for Cronbach's alpha in Study 1A	114
Table 4.7 Rank order of items by difficulty across conditions	115
Table 4.8 Item discriminations across conditions	116
Table 4.9 Demographic Details: Study 1B	118
Table 4.10 Number of Visits to Interaction Space by Condition	120
Table 4.11 Average Duration of Whole Fixations on Interaction Space by	

Condition	121
Table 4.12 Proportion of Whole Fixations on Interaction Space by Condition	122
Table 4.13 Demographic and Performance Details: Study 2	124
Table 4.14 Difficulty and Discrimination Indices for Study 2.....	124
Table 4.15 Item Performance by Task	125
Table 4.16 Mean Time-on-Task(s) by Total Item Score.....	126
Table 4.17 Mean Time-on-Phase (per task) correlated with Item Performance (all parts).....	128
Table 4.18 Relationship between Number of Simulations run and Item Performance (all parts).....	129
Table 4.19 Mean Number of Visits to AOIs in Orientation Phase	131
Table 4.20 Mean Times-To-First-Fixation (ms) on Relevant AOIs (Output Phase)	132
Table 4.21 Mean Times-To-First-Fixation (ms) on Relevant Areas by Item Performance.....	133
Table 4.22 Time-To-First-Fixation on Relevant Information (ms) in Output Phase: Comparisons of Partial/No Credit (Group 1) and Full Credit (Group 2)	133
Table 4.23 Mean Number of Fixations on Relevant Information (Output Phase)	134
Table 4.24 Number of Fixations on Relevant Information in Output Phase: Comparisons of Partial/No Credit (Group 1) and Full Credit (Group 2)	135
Table 4.25 Mean Proportion of Fixations on Relevant and Irrelevant Areas (Output Phase)	137
Table 4.26 Proportion of Fixations on Relevant Information in Output Phase: Comparisons of Partial/No Credit (Group 1) and Full Credit (Group 2)	138
Table 4.27 Profile of Interviewees for Study 3	140
Table A1 Types of Cognitive Load (Chandler & Sweller, 1991; Mayer, 2005)	199
Table B1 Explain Phenomena Scientifically	203
Table B2 Evaluate and Design Scientific Enquiry	203

Table B3 Interpret Data and Evidence Scientifically	204
Table B4 Categories describing the scientific literacy items constructed for the PISA 2015 cycle.....	205

List of Figures

Figure 1.1 Research Focus	11
Figure 2.1 A possible classification system for technology-based items (e.g. Bennett, 2015; Russell & Moncaleano, 2019)	17
Figure 2.2 Interaction Spaces available on the Surpass Platform (BTL, 2018)	18
Figure 2.3 Multimedia SJT and Response Options (Reprinted from MacCann et al., 2016)	22
Figure 2.4 Examples of Figural Response (FR) items (Professional Testing, 2018) ..	23
Figure 2.5 Sample interface from <i>Knowledge Maps</i> (Reprinted from Ho et al., 2018)	25
Figure 2.6a Sources first (Reprinted from Arslan et al., 2019)	26
Figure 2.6b Targets first (Reprinted from Arslan et al., 2019).....	27
Figure 2.6c Swapped content (Reprinted from Arslan et al., 2019)	27
Figure 2.6d No problem statement with instructions (Reprinted from Arslan et al., 2019)	27
Figure 2.7 Image, animation and simulation conditions (Reprinted from Quellmalz et al., 2013)	30
Figure 2.8 Classification System for Technology-Based Items	32
Figure 2.9 Cognitive Theory of Multimedia Learning (adapted from Mayer, 2009, p. 61)	38
Figure 2.10 Intrinsic, Extraneous and Germane Assessment Load for experts and novices for a difficult task (Reprinted from Kirschner et al., 2016, p. 23)	48
Figure 2.11 Evidence Centred Design (ECD) conceptual assessment framework (modified from Hao & Mislevy, 2018)	53
Figure 3.1 Conceptual Framework underlying the current study.....	62
Figure 3.2 Outline of Study 1 (A, B), Study 2 and Study 3	67
Figure 3.3 Convergent parallel design for mixed methods research	69

Figure 3.4 2x3 Mixed Factorial Design	70
Figure 3.5 Participant numbers and contributions to data collection (Study 1, Study 3).....	73
Figure 3.6 Participant numbers and contributions to data collection (Study 2, Study 3).....	73
Figure 3.7 ‘Meteoroids and Craters’ (Questions 2-3, PISA 2015 from OECD, 2020) ..	75
Figure 3.8 ‘Bird Migration’ – Original PISA 2015 stimulus, stimulus for static condition, stimulus for animated condition	79
Figure 3.9 Item from ‘Groundwater Extraction and Earthquakes’ unit (Static, Dynamic)	80
Figure 3.10 ‘Running in Hot Weather’ (Question 2, PISA 2015 from OECD, 2020)	81
Figure 3.11 Tobii Pro Fusion Eye-Tracker	85
Figure 3.12 Placement of AOIs for an SR item in Study 1B (Static, Dynamic)	92
Figure 3.13 Placement of AOIs for ‘Orientation’ phase of items in Study 2.....	94
Figure 3.14 Placement of AOIs for ‘Output’ phase of items in Study 2	95
Figure 3.15 Visual representation of calibration process	101
Figure 4.1 Summary of Data Collected (by Study).....	107
Figure 4.2 Percentages of students who agreed/disagreed with various statements about their interest in different science topics.....	110
Figure 4.3 Comparison of item discriminations across conditions.....	117
Figure 4.4 Time-on-Task for Orientation and Output phases.....	127
Figure 4.5 Heat map for the ‘Orientation’ phase of Task 3 ($n=24$).....	130
Figure 4.6 Relative count of fixations for Task 3.....	136
Figure 4.8 Thematic frame representing principal themes and subthemes	142
Figure 5.1 Conclusions	166
Figure B1 ‘Groundwater extraction and Earthquakes’ (PISA 2015)	204
Figure C1 Item M1 (Static, Dynamic)	206

Figure C2 Item M2 (Static, Dynamic)	207
Figure C3 Items M3, M4 (Static, Dynamic)	208
Figure C4 Item F1 (Static, Dynamic – Identical items across conditions)	209
Figure C5 Item F2 (Static, Dynamic – Identical items across conditions)	210
Figure C6 Item F3 (Static, Dynamic – Identical items across conditions)	210
Figure C7 Item P1 (Static, Dynamic)	211
Figure C8 Item P2 (Static, Dynamic)	212
Figure C9 Item P3 (Static, Dynamic)	213
Figure C10 Item P4 (Static, Dynamic)	214
Figure C11 Item G1 (Static, Dynamic)	215
Figure C12 Item G2 (Static, Dynamic)	216
Figure C13 Item G3 (Static, Dynamic)	217
Figure C14 Item G4 (Static, Dynamic)	218
Figure E1 Explanatory Text/Practice Task	220
Figure E2 Task 1	221
Figure E3 Task 2	221
Figure E4 Task 3.....	222
Figure E5 Task 4	222
Figure E6 Task 5	223
Figure F1 Good vs Poor Eye Movement Accuracy and Precision (adapted from Dalyrmple et al., 2018)	224
Figure F2 Tobii Pro Fusion	225

List of Acronyms and Abbreviations

AERA	American Educational Research Association
AOI	Area of Interest
APA	American Psychological Association
CR	Constructed Response
CTMA	Cognitive Theory of Multimedia Assessment
CTML	Cognitive Theory of Multimedia Learning
DES	Department of Education and Skills
FR	Figural Response
M	Mean
Md	Median
NCCA	National Council for Curriculum and Assessment
OECD	Organisation for Economic Co-operation and Development
PISA	Programme for International Student Assessments
SD	Standard Deviation
SEC	State Examination Commission
SR	Selected Response
TBA	Technology Based Assessments
TEI	Technology Enhanced Items
TOI	Time of Interest

Acknowledgements

Firstly, I would like to thank my supervisor, Michael O’Leary: a teacher, scholar and mentor who I greatly admire and owe so much to. I feel hugely privileged to have worked with Michael and I am a better person for having his guidance on this journey. Michael’s unconditional kindness, unfailing work ethic, impeccable use of emojis, and words of encouragement gave me the confidence I needed to complete this thesis and take on new challenges. I sincerely appreciate all the time, effort and energy that he has invested in me over the past three and a half years. I promise Michael that I will put your philosophy into action and ‘pay it forward’ whenever I can in the future. Thank you.

My sincere thanks also go to my secondary supervisor, Darina, who, though very busy herself, was always available to answer my bizarre questions that sometimes bore no relation to my thesis topic. She took my rambling voice notes in her stride and reassured me whenever I thought the PhD was ‘going wrong’. I am lucky that she was my supervisor during this process and I am even luckier now that I can call her a friend.

I wish to thank my third supervisor, Mark Brown, whose feedback on my final draft gave me the confidence to realise that I was ready to (finally!) submit. I also want to acknowledge the contribution of my panel member, Zita Lysaght, who was always willing to offer a friendly smile, amusing anecdote or insightful advice. Thank you both.

I am very grateful for the financial support I received from the Irish Research Council as it allowed me to work full-time on my thesis and acquire the very expensive eye-tracking machine. I would also like to thank Prometric for their initial funding as well as Linda Waters and Li-Ann Kuan for supporting my involvement in other projects.

My sincere thanks also go to Vasiliki and Conor, my upstairs/downstairs office mates in Moville who are now my friends for life. Despite the best efforts of the seagulls, you both made the office a wonderful place to be with your dog pictures and random facts. I also wish to thank the other members of CARPE for their support over the years, particularly Anastasios whose serene approach to his own PhD was a model for me to follow (or at least attempt to). I am grateful to all my other colleagues in Dublin City University who supported me at different stages of my research: Deirdre Butler, Margaret Leahy, Anne Looney, Brian MacCraith, Elaine McDonald, Cillian Murphy and Maeve Power.

The six schools who facilitated my research despite the huge challenges they were already dealing with because of a global pandemic will forever have my gratitude. Without them, this project would never have succeeded and I would still be stress baking cookies. Paul Behan of the National Council for Curriculum and Assessment also has my sincere thanks for his support during this time as well.

To the Board of Management, my principal Sinéad and all my colleagues at Belmayne ETNS: thank you for all your support over the past three and a half years and for always welcoming me back for visits. Thank you all for the laughter, patience and mystery trips (even if I did nearly make us miss our flight to Madrid!).

I am eternally grateful to all my friends. To Aisling, Susan, Eileen and Louise who have been with me since the beginning: thank you for all your support and those ‘Friday night zooms’. I am also very appreciative of Aisling’s proofreading skills! I am lucky to

have such enduring friends. I am also grateful to Denise, Johanne and Viv for always picking up the phone or going on walks with me. They kept me grounded and sane at a time when it was easy to feel neither.

I want to say a special thanks to Denis who first sent me the advertisement for the PhD position in CARPE and told me to ‘have a think’ about it. He has always been the very definition of a true friend and I thank him for all the laughs, encouragement and support over the past decade or so. I am also very grateful to Tom (and Katya) as well for letting me borrow him as necessary and for always reminding me that their home was mine as well.

To my sisters, thank you for never looking too bored when conversations inevitably turned to ‘the PhD’. To Deirdre, thank you for providing me with a heavy duty printer and endless treats. To Elaine, thank you for the Sunday ‘coffee and supermarket’ trips and for telling me that I would always get here in the end. To Clare, thank you for always accepting my multiple phonecalls, for helping me to calmly solve all my problems (both real and imaginary) and for your superior editing skills. Thanks as well to Alex, Elise, Luke, Rían and Cillian – you were all the best distractions ever.

Finally, it remains for me to sincerely thank my parents, Francis and Mary. While I cannot exactly recommend doing a PhD during a pandemic, you both made it a lot easier! Thank you for offering up the spare bedroom as an office and for always checking in with me even when it may have been safer to leave me alone. I never would have come so far without your continuous belief in me and your reminders that I can achieve anything with common sense, hard work and perseverance. This thesis would not have been possible without your support and sacrifices and I dedicate it entirely to you both. You will always be my first, and best, teachers.

Abstract

The Impact of Test Items Incorporating Multimedia Stimuli on the Performance and Attentional Behaviour of Test-Takers

Paula Lehane

Many countries are now deploying online testing solutions for their terminal post-primary exams e.g. Ireland, New Zealand. These Technology-Based Assessments (TBAs) use items that employ a broad array of interactive, dynamic or static stimuli e.g. simulations, animations, text-image. Although it is assumed that these features can make TBAs more authentic and effective, their impact on test-taker performance and behaviour has yet to be fully clarified.

This research investigated the extent to which the use of different multimedia stimuli can affect test-taker performance and behaviour using a mixed methods approach. Guided by four main research questions, an experiment was conducted with 251 Irish post-primary students using an animated and text-image version of the same TBA of scientific literacy. Eye movement and interview data were also collected from subsets of these students ($n=32$ and $n=12$ respectively) to determine how differing multimedia stimuli can affect test-taker attentional behaviour. A second study involving 24 test-takers completing a series of simulation-type items was also undertaken. Eye movement, interview and test-score data were gathered to provide insight into test-taker engagement with these items.

The results indicated that, overall, there was no significant difference in test-taker performance when identical items used animated or text-image stimuli. However, items with dynamic stimuli often had higher discrimination indices indicating that these items were better at distinguishing between those with high and low levels of knowledge. Eye movement data also revealed that dynamic item stimuli encouraged longer average fixation durations on the response area of an item. An examination of the data relating to test-taker performance and behaviour for simulation-type items found that test-takers developed more efficient search strategies as their familiarity with this item type increased. The data also showed that there was a weak to moderate relationship between task performance and time-to-first-fixation on relevant information. The implications of these and other findings, as well as recommendations for policy, practice, and future research are discussed in the final chapter of this thesis.

Chapter 1

Introduction

1.1 Research Topic and Problem

Given the widespread proliferation of digital technology in everyday life, it seems inevitable that all assessments¹, regardless of discipline or sector, will be administered through this medium in the future (Bakia et al., 2011). Therefore, it is essential that test-developers know how to design reliable and appropriate items² for all Technology-Based Assessments (TBAs), including digital tests³. While paper-based tests are largely restricted to traditional multiple choice, short answer or essay questions, the possibilities for items in TBAs are potentially limitless. For example, items in TBAs can include many unique multimedia features and objects such as high-resolution images, animations and even simulations (Bryant, 2017). Items in TBAs can also be highly interactive, often requiring participants to respond to test items in complex ways that were not previously seen in paper-based testing contexts or in earlier versions of TBAs. To provide some much needed knowledge on these medium-unique features, this thesis explored the design of test items for TBAs. More specifically, it examined if the use of different multimedia objects (e.g. images, animations) or question types (e.g. simulations) can impact test-taker performance and attentional behaviour in digital tests.

In particular, this thesis explored the design and use of items for TBAs in educational settings. In 2015, 58 of the 72 countries who participated in the Programme for International Student Assessment (PISA) administered the technology-based version of these tests (Organisation for Economic Co-Operation and Development [OECD], 2016a). Data from these international large-scale assessments have come to dominate education-policy discussions and decisions. Countries often use their students'

¹ When applied to educational contexts, *assessment* can be considered a procedure for making inferences about learning (Cronbach, 1971). This procedure involves the collection, synthesis, interpretation and use of data to answer questions, solve problems or facilitate decision making regarding student progress (Russell & Airasian, 2012). Technology-based assessments involve the use of electronic software and computerised devices such as personal computers, laptops, and tablets to engage in this process (Mayrath et al., 2012).

² Within the field of assessment, the term *test item* is used to refer to the questions used in the test (Russell & Airasian, 2012). Each item includes a stem 'which presents the problem or question to the student' (Russell & Airasian, 2012, p. 146). The students must then respond to the test item by selecting an answer from a set of options or constructing their own responses using text or diagrams.

³ According to Russell and Airasian (2012, p. 11), a *test* is a 'formal, systematic procedure used to gather information about students' achievement or other cognitive skills'. It involves the administration of a set of questions, the completion of which will provide some measure about the test-taker's performance. For the purposes of this thesis, *digital tests* involve the use of technology in the administration and delivery of the assigned test questions.

performance in this test as a standard by which to judge their education systems. For many countries, including Ireland, downward fluctuations in the performance of their students were noted in 2015, with some attributing these variations to the interactive, multimedia-heavy items contained in the TBA used (e.g. Jerrim et al., 2018). Such items may also reduce the comparability of scores across and within PISA cycles (Jerrim et al., 2018). Although Bryant (2017) argued that these items can make TBAs more authentic and can measure a greater array of knowledge and skills, the impact of these medium-unique items on test-taker performance has yet to be fully investigated. Preliminary research involving TBAs that use multimedia stimuli like animations have been found to alter the information processing and attentional allocation behaviours of test-takers which could potentially influence what test score an individual achieves (e.g. Malone & Brünken, 2013). Hence, it is possible that the inclusion of multimedia stimuli in TBAs may modify the knowledge or skill being assessed or introduce some unknown facet that could result in an erroneous judgement of competency (Vorstenbosch et al., 2014). Poorly designed test items therefore, could have serious ramifications for test validity.

Despite these concerns, education systems around the world are now attempting to follow PISA's example and devise their own TBAs. For example, examinations in New Zealand's National Certificate of Educational Achievement (NCEA) will be done entirely through computers by December 2021 (New Zealand Qualification Authority [NZQA], 2018). Similarly, Ireland developed a computer-based exam for the newest Leaving Certificate subject of Computer Science that was first deployed in May 2021 (State Examination Commission [SEC], 2021). As achievement in this exam contributes to a student's overall success in the Leaving Certificate, which can then mediate the future courses of study available to them in further and higher education, it is essential that this exam, and any future ones, are appropriately designed (Lehane, 2019).

1.2 Significance of the Study

TBAs do not make sole use of Multiple-Choice Questions (MCQs) or basic text-entry constructed-response items. Instead, they use a range of items that consist of tasks which employ a broader array of stimuli (e.g. audio, video, interactive graphs) and response mechanics (e.g. drawing, drag-and-drop) (Oranje et al., 2017). These can take the form of:

‘single stem questions (i.e. a stimulus that is followed by a single question), simulation based tasks (e.g. a virtual experiment that has to be conducted including setting up the experiment, running the experiment, recording findings and synthesising results), scenario-based tasks..., and sandbox type tasks’

(Oranje et al., 2017, p. 39)

An insufficient understanding of the optimal design of these items for TBAs could jeopardise the validity of educational assessments. According to the latest *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014), validity ‘refers to the degree to which evidence and theory support the interpretations of test scores for the proposed uses of tests’ (p. 11). It is the most fundamental consideration in developing and evaluating tests. To safeguard the validity of a test, Messick (1994) asserts that a clear evidence structure should be in place. Therefore, test items that involve complex stimuli or actions, such as those described by Oranje et al. (2017), require a systematic evaluation to determine whether or not the actions undertaken by test-takers while performing an item are actually consistent with the intended claims about the test-taker made based on their test score (Lane, 2017). Research to better understand what evidence items with different stimuli and response mechanisms can provide is required. This study aimed to address this validity issue through its investigation of different item types (e.g. MCQs, simulations) with varying stimuli (e.g. animations, images) using eye movement data and cognitive think-alouds.

1.2.1 Building ‘good’ items: Understanding their features and characteristics

The primary purpose of any item in an assessment, irrespective of its capacity to provide advanced response actions or stimuli, is to ‘collect evidence of the test-taker’s development of the targeted knowledge, skill or ability’ (Russell, 2016, p. 21). The targeted knowledge, skill or ability may also be referred to as a construct⁴ within

⁴ According to Fulcher and Davidson (2008) a ‘construct’ is any theory, hypothesis or model that attempts to explain an observed phenomenon. Assessments aim to provide operational definitions of constructs to support the measurement process. Defining constructs can be difficult and approaches can vary by field and discipline. For example, Bachman (1990) notes that language tests use three approaches for defining constructs: 1) ability-focused, 2) task-focused, and 3) interaction-focused (constructs can only be measured within particular contexts).

literature. Test items should contribute to an assessment's overall ability to '*fully* represent all the knowledge, skills, and abilities inherent in the construct being measured' (Sireci & Zenisky, 2006, p. 300). This is called construct representation. The most common types of items used to measure constructs in tests are MCQs and open-ended questions, where test-takers 'construct' a response using text or graphics. Messick (1988) noted that 'tests are imperfect measures of constructs because they either leave out something that should be included . . . or else include something that should be left out, or both' (p 34). Leaving something out causes construct underrepresentation. Longstanding complaints that traditional, text-based MCQ items are limited in what they measure (as outlined by Scully, 2017) contributed to the development of technology-based items. These items include more complex stimuli in the forms of images and animations but also require more complex item response actions from test-takers. In the 2015 PISA TBA, some test items required students to run simulations involving multiple actions (moving sliders, selecting different variables) by the test-taker in order to better measure test-takers' proficiencies in running scientific enquiries and interpreting data (OECD, 2016b). Others involved students 'dragging and dropping' their answers into an ordered list instead of just selecting them. In the US, standardised assessments for primary and post-primary students, as well as those in the field of credentialing, have also begun to include constructed response items that allow for more complex graphic responses to be included (Masters & Gushta, 2018). These ways of responding to and representing items would not have been available in paper-based tests or in earlier versions of TBAs. Some argue that items requiring complex response actions that go beyond the simple or sole selection of an answer facilitate better construct representation and are better test items (Dolan et al., 2011).

However, it could also be argued that by attempting to address concerns over construct underrepresentation, less attention has been paid to the risks of these technology-based items; specifically, that they include something that should be left out. For example, test-taker unfamiliarity with simulations or poor typing proficiency could have a negative impact on test-taker performance. This construct-irrelevant variance⁵

⁵ Messick (1988, p. 34) defined construct-irrelevant variance as 'excess reliable variance that is irrelevant to the interpreted construct'. The variance is systematic in nature (Downing & Halydyna, 2006) and can be caused by poor test design. For example, increased item difficulty caused by poorly written or designed items is an example of construct irrelevant variance. Poorly designed test questions add artificial difficulty to test scores which then negatively impacts on the validity of test scores as accurate interpretations are not possible.

occurs when factors unrelated to the target construct are measured which then affects test-scores in a systematic manner (although, not necessarily for all test-takers). Although construct-irrelevant variance is a significant risk to the validity of assessments, current literature discussing the design, development and definitions surrounding technology-based items is more likely to address construct representation. To fully understand how items can support construct representation and how construct-irrelevant variance can be avoided, research on item formats and types should aim to 'examine item design features at a fine-grained level rather than making sweeping assumptions for a given type' (Moon et al., 2019, p. 61). When in-depth research on item types are combined with appropriate cognitive theory, empirically grounded item writing guidelines can be established. This research study attempted to realise this aim in relation to the use of multimedia item stimuli and items involving simulations.

1.2.2 Selecting the right stimulus: Images, animations and simulations

Various types of multimedia elements can now be found in educational TBAs. 'Multimedia' refers to the combination of text with other media elements such as images, animations or simulations to communicate meaning and information (Jordan, 1998). The addition of multimedia objects to an item can greatly affect test-taker performance. For example, work by Lindner et al. (2017a) found that the addition of representational pictures⁶, in the form of illustrations, to text-based items improved student performance, accelerated item processing and reduced rapid guessing behaviours in testing contexts. Therefore, it is hardly surprising that there is now a growing discussion around the use of animations, the dynamic version of illustrations, for test items (e.g. Tuzinski, 2013). According to Mayer and Moreno (2002), animations depict a 'simulated motion picture... [showing] movement of a drawn (or simulated) object' (p. 88)⁷. Animations can present test-takers with a more realistic picture of a given situation and can communicate more complex concepts and information that cannot be quickly communicated with text (Tuzinski, 2013). Furthermore, when replacing text-based stimuli, animations can reduce certain construct-irrelevant variance related to reading or language proficiency. This is

⁶ Representational pictures 'visualise the item-stem text but do not add any other solution-relevant information' (Lindner et al., 2017a, p. 482).

⁷ Based on this definition, animations require the following: (a) a pictorial representation, (b) a depiction of movement, and (c) objects are artificially created through drawing or some other simulation method (Mayer & Moreno, 2002).

particularly relevant to the measurement of non-cognitive constructs like interpersonal skills or motivation. A recent study by Karakolidis et al. (2021) compared test-taker performance of native ($n=51$) and non-native ($n=66$) English speakers using an animated and text-based version of the same situational judgement test⁸ related to the 'practical knowledge' (a non-cognitive construct) of those involved in the teaching profession. The variance attributed to construct-irrelevant factors like native language and reading comprehension in English was lower by 9.4% in the animated version of the test.

Animations may have a facilitative effect on test-taker performance. While Karakolidis et al.'s (2021) work shows the value of animations in reducing construct-irrelevant variance for certain populations, concerns about the inclusion of animations, and indeed images, in tests for other populations exists. For example, Wu et al. (2010) used a comparative experimental design where participants were stratified into three categories (low, medium and high) depending on their level of prior knowledge for the test topic (Earth Science). The participants then completed a test that used animated or static stimuli. Participants with low levels of prior knowledge performed better in the animated condition (Cohen's $d=0.7$) while those with higher levels of prior knowledge performed better when static pictures were used (Cohen's $d=0.7$). Malone and Brünken (2013) also uncovered similar findings when comparing the scores of expert and novice drivers on items that included static and animated stimuli. This interaction effect between stimulus type (image, animation) and expertise level could have significant implications for the selection of item stimuli in tests. The current research study compared the use of animated and static stimuli on test-taker performance using previous performance on school-based tests as a measure of science proficiency. The simultaneous collection of process data in the form of eye-movements from some participants in this research provided additional insights on the behavioural processes associated with test-takers' performance according to stimuli type and proficiency that were not available in previous studies such as those of Wu et al. (2010) and Malone and Brünken (2013).

The inclusion of simulations in test items has also become more commonplace with educational TBAs. Simulations are *interactive* forms of multimedia objects whereby test-takers can 'produce' an imitation of a real world scenario (Levy, 2012). When used

⁸ SJTs simulate realistic job-related situations where test-takers are presented with a described scenario and asked to choose an appropriate reaction or response from a list of different options (Lievens & Sackett, 2006).

in assessment contexts, simulations assess a test taker's proficiency on the basis of their interactions with the virtual environment and their ability to use any information they have generated to answer other items or tasks (Baker & Clarke-Midura, 2013). While use of these items are growing in popularity (e.g. OECD, 2016a), some commentators have noted that their introduction to educational TBAs has been somewhat rushed from a practical and psychometric perspective (e.g. Shiel et al., 2016; Lee et al., 2019). There appears to be a lack of understanding as to how test-takers engage with these multimedia objects in an assessment context (Teig et al., 2020) and the most effective ways to score such items (Lee et al., 2019). Research is currently trying to address such shortcomings in order to better understand how to best use these interactive multimedia objects. For example, Lee et al. (2019) explored test-takers' interactions with simulations using a response-time perspective. This study yielded valuable information on the timing aspects of test-takers' behaviours with simulation-type items. However, they noted that further information on test-takers' interactions with simulations on a range of other behaviour-based variables (e.g. mouse clicks) was needed as a matter of urgency to assist in the more effective design, assembly and scoring of items that included simulations. The current study also attempted to address this.

1.2.3 Leveraging process data to inform item design and validity claims

When making any claims about what a test and its items are supposed to assess, there should be clear evidence to support these claims. For example, the newly developed simulation items for the scientific literacy domain in PISA 2015 (OECD, 2016b) claim to assess students' abilities to conduct scientific inquiries and interpret data. Students are required to run a series of simulations to find the data needed to answer the given question. It is assumed that students who provide the correct answer can conduct scientific inquiries and interpret data. Using this reasoning, it is then expected that *all* those who provided an incorrect answer are assumed to be *unable* to conduct scientific inquiries and interpret data. However, in a complex task such as this, the inferential distance between the claims of what the item assesses and the subsequent evidence provided through test-taker scores can be quite large (Oranje et al., 2017). For constructs that are complex and focussed on the application of procedures, sequences of actions (e.g. modifying variables to run multiple simulations) may represent important evidence about the construct rather than just the final answer. Therefore, information on *how* test-

takers engage with certain items can allow for the development of more accurate test claims. This information could then be used to identify different levels or degrees of test-taker knowledge. Obtaining this information is best done through the use of process data.

Interviews, behavioural assessments, and self-reporting measurements are the most common tools used to make inferences about how people process information from multimedia stimuli in learning and assessment contexts (Alemdag & Calgitay, 2018). Unfortunately, self-report measures can be particularly unreliable and cannot capture those processes of cognition that people are unaware of. To overcome the limitations of self-report measures, the *Standards* (AERA et al., 2014), recommend the use of response process data which is any 'evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees' (p. 15). Eye-tracking technology can be used to explore how test-takers process information in TBAs involving different multimedia tasks and activities. Eye-tracking refers to 'a set of technologies that make it possible to establish the eye gaze of an individual' (Navarro et al., 2015, p. 2237). By using infrared beams that are reflected onto an individual's pupils and then recorded, Hyöna (2010) noted that eye-tracking technology can allow researchers to identify what is attended to first in a presentation, and for how long, along with other information related to the attentional allocation processes of humans. The current study used eye-movement data to better understand the attentional behaviour of test-takers when answering questions involving different multimedia stimuli and response actions.

1.3 Study Origins and Researcher Positionality

Developments in digital assessment can offer substantial value to society. For example, the Road Safety Authority (RSA) of Ireland developed a digital driver theory test that all drivers must pass before gaining a provisional driving license, which they claim has contributed to the significant decline in fatalities on Irish roads since 2008 (RSA, 2018). Yet, a set of clear guidelines relating to the optimal design of digital assessments has yet to be identified within any field or discipline (Bryant, 2017). This is concerning given the how commonplace digital assessments and TBAs are becoming within credentialing and educational settings. For example, in 2019 digital versions of standardised English Reading and Mathematics tests were recently deployed in Irish primary schools for the first time by the Educational Research Centre (ERC; 2018). It is

important to note that the researcher of this study was an active agent in the Irish education system which afforded them a deep knowledge and understanding of the dynamic and ever-changing field of educational assessment. Therefore, as a researcher's expertise, beliefs, values and experiences can affect the knowledge that they construct from research (Sikes, 2004), it is necessary to consider and acknowledge the researcher's own positionality on issues related to their field of study. This ensures that there is a better understanding of the 'assumptions [they hold] which inform their sense of the world' and how that could have 'implications for their research' (Washington et al., 2005, p. 21).

Prior to beginning this study in 2018, the researcher was a full-time primary school teacher for seven years. During this time, the researcher noted that digital assessments were being distributed more regularly within primary and post-primary schools, particularly for the diagnosis of special educational needs. When administering these tests, the researcher noted that students were often distracted by timers, illegible fonts or unnecessary animations and images. They also became frustrated when they could not navigate the interface or respond to test items. This often had a negative impact on the students' test performance which resulted in the researcher questioning the appropriateness of the judgements being made about these students. Consequently, the researcher felt compelled to further examine the optimal design of digital assessments for primary and post-primary students. The researcher's position as an 'insider' within the Irish education system was advantageous for identifying and approaching respondents as well as designing appropriate research materials as outlined by Merriam et al. (2005). However, this position and the researcher's experiences may have also been a source of bias. For example, 'insiders' are often at risk of being 'inherently biased' according to Merriam et al. (2005, p. 411) e.g. previous experiences making them sceptical on the use of digital assessments. Reflection on the researcher's potential biases and how to best address them informed her research approach and design decisions such as the use of data triangulation.

1.4 Scope of the Study

This research study examined the impact of item design on test-taker performance and attentional behaviour in TBAs using test items that measure post-primary school students' scientific literacy. Scientific literacy, as defined by the OECD (2016b), involves

‘the ability to engage with science related issues, and with the ideas of science, as a reflective citizen’ requiring an ability to ‘explain phenomena scientifically, to evaluate and design scientific enquiry, and to interpret data and evidence scientifically’ (p. 50). The items selected for use in this study were deployed in the development and administration phases of the 2015 PISA cycle where scientific literacy was the major domain tested (OECD, 2016a). However, it is important to acknowledge that the current study is not an attempt to measure or conceptualise scientific literacy. This has already been done by those who developed and designed PISA. Instead, the scientific literacy items used in this study acted as a vehicle to determine if certain item features can contribute to test-taker performance and behaviour.

The aim of this study was to determine if an item’s stimulus can influence the performance or attentional behaviour of test-takers in a TBA. In relation to item stimuli, the current study compared the possible impact of two types of item stimuli on test-taker attentional behaviour and performance – animated stimuli and text-image stimuli. Static images with text and basic animations appear to be the most common stimuli in educational tests for primary and post-primary school-aged children worldwide (e.g. PISA, Drumcondra Primary Tests of Reading and Mathematics, Smarter Balanced Assessments etc.). As a result, investigations in this research were focussed on these forms of item stimuli. Similarly, select-only and constructed response type questions are the dominant item types used in these tests as well (Bryant, 2017). However, more complex response actions are now possible e.g. ‘drag-and-drop’ items. They also include the use of simulation-type items which can include interactive multimedia stimuli (OECD, 2018). The scope of this research study was therefore restricted to exploring the impact of select-only, constructed response and a small selection of these ‘new’ types of items. Figure 1.1 outlines the focus of this research highlighting the specific areas of interest (Item Stimuli, Response Actions) along with the elements of test-taker engagement they were investigated under (attention, test scores).

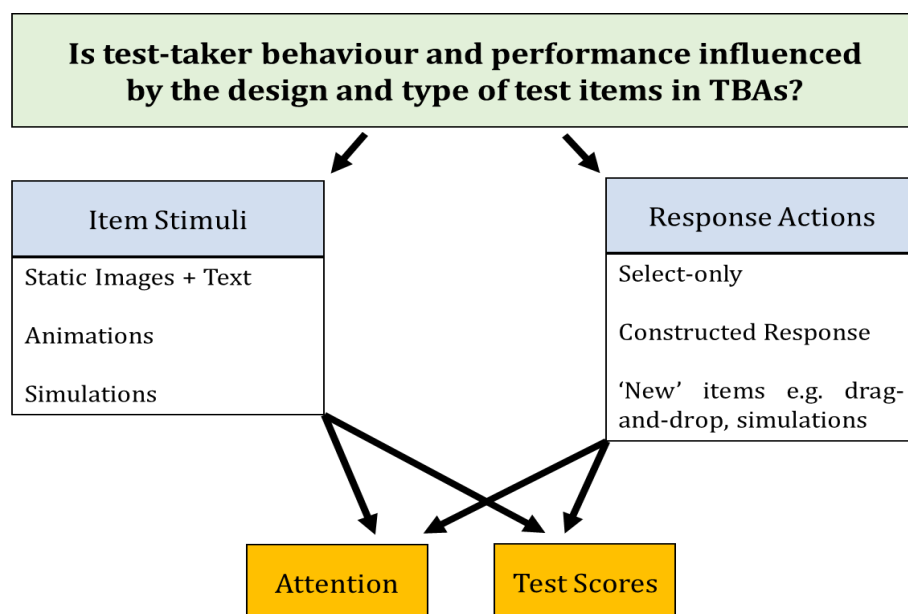


Figure 1.1 Research Focus

To determine if certain item features can influence test-taker performance, data from test-scores, eye-movements and think-aloud protocols were collected in this study to allow for the triangulation of information. In particular, eye movement data were included in the current study as the main source of process data to validate test score meaning in line with Oranje et al.'s (2017) recommendations. While log data⁹ and response times were also available for use, eye movement data were considered to be the most appropriate form of process data for this study as it traced the location of an individual's gaze throughout the test-taking process. Eye gaze suggests where visual attention and cognitive processing is focussed (Just & Carpenter, 1980). Therefore, the eye movement data collected in this study provided insights as to how certain item features can influence test-taker attention. When combined with test score data and information from think-aloud protocols, a better understanding of the cognitive implications of item formats can be obtained.

1.5 Organisation of the Thesis

This introductory chapter represents the first of five chapters. The next chapter explores the literature and previously conducted research regarding technology-based

⁹ Log data records all events in an online environment e.g. what answer options were selected or de-selected, how a slider was moved etc., (Oranje et al., 2017).

test items. This chapter will deliberate what issues should be considered when designing items in TBAs, particularly in relation to the use of multimedia stimuli and different response actions. These discussions on item design will be grounded in appropriate psychological learning theory and will explore if the application of such theories to assessment contexts is appropriate and what type of research should be conducted to address such questions. This critical examination of relevant literature and research will inform the research aims of the study. Chapter 3 will outline the methods employed to address the research questions and the procedures undertaken. Chapter 4 presents the research findings which will be critically reviewed in Chapter 5. This final chapter will also provide recommendations for policy, practice and future research within the field.

Chapter 2

Literature Review

2.1 Introduction

It has been suggested that an insufficient understanding about the optimal design of items for Technology-Based Assessments (TBAs) could jeopardise the appropriateness of the inferences drawn from an individual's performance on them (e.g. Bryant, 2017; Embretson, 2016). To determine the veracity of this assertion, this chapter critically analyses previous research and literature on item design for TBAs. The chapter will begin with an overview of TBAs and an attempt to classify the different types of test items associated with them. In particular, the impact of including multimedia objects (e.g. animations, images, simulations) in these test items will be considered. The need for a cognitive theory of multimedia assessment to guide the design of items that use such multimedia objects will be deliberated with reference to relevant psychological theory. From this, a discussion on how to best advance the field of TBAs will be undertaken, with particular reference to the central role process data should play in understanding the value of different item designs. The chapter will conclude with a summary of the key issues within the field and will identify the major strengths and shortcomings in available research literature on the impact of item design decisions in TBAs.

2.2 Scope of the Literature Review

A comprehensive search strategy was used to obtain studies relevant to the topic of item design. Initial scoping searches included the exploration of a range of electronic index databases, including PsychINFO, PubMed, ERIC, EBSCO, Social Sciences Citation Index and Web of Science. Databases outside the discipline of education were explored given the relevance of the topic to other contexts (e.g. credentialing). Non-indexed databases including Dissertation Abstracts, Digital Dissertations and ScienceDirect were also used along with Google Scholar. Key terms, and their synonyms, employed in the initial scoping searches included 'technology based assessments/ computer based assessments', 'technology enhanced items', 'alternative items', 'innovative items', 'item format', 'multimedia' and 'response option'. The Boolean operators of 'AND' and 'OR' with

the terms 'efficacy', 'design', 'validity', and 'utility' identified research that explored the relative value of different item design decisions. No deliberate time frame was applied to the studies returned from this search to ensure that all possible information on this issue was available. The reference lists of selected studies were also screened for other potentially relevant papers. The studies presented in this literature review were selected solely based on their relevance to the research topic. Consequently, findings from peer-reviewed journals within the realm of educational assessment, human-computer interaction and cognitive psychology, along with the grey literature of unpublished manuscripts and technical reports from testing organisations, were all considered.

2.3 Technology Based Assessments (TBAs)

The evolution of TBAs in education can be conceptualised in terms of three 'generations' as outlined by Bennett (2015). First-generation TBAs involve the delivery of traditional assessments via computers. O' Leary et al. (2018) acknowledged that this is a basic, one-time event that takes limited advantage of technology. Ireland's recent development of a digital version of the paper-based standardised tests used in Irish primary schools is one such example of this (ERC, 2018). Second-generation TBAs aim to maximise efficiency and improve quality by deploying innovative item formats and automated scoring procedures. There are also efforts to use technology to assess 'hard to measure' skills, that could not be previously assessed using paper and pencil tests. In this way, PISA's (OECD, 2018) recent TBAs can be classified as second-generation assessments. However, when devising digital assessments, the aim should be to harness technology in a way that truly enhances the assessment process, by expanding the possibilities of what can be assessed, or the range of inferences that can be made from assessment results. It is only from this that third-generation TBAs can emerge. Items for this generation of TBAs are designed according to general cognitive principles and theory-based domain models. Bennett (2015) noted that they can be identified through their use of 'complex simulations and other interactive performance tasks' (p. 372). Third-generation assessments incorporate multiple sources of information to create coherent models to provide more accurate information about test-takers. O' Leary et al. (2018) claimed that third-generation TBAs are 'models of effective pedagogical practice' that go beyond evolution to revolution (p. 161). Some of the items involving simulations that were contained in PISA 2015 (OECD, 2017) alluded to such features. The building

blocks for third-generation assessments are beginning to emerge. Although Bryant (2017) argued that such resources can make TBAs more authentic than their paper-based counterparts and can measure a greater array of knowledge and skills, their inclusion in test items for TBAs presents a number of issues.

2.4 Technology-Based Items

Due to the rapid and ever-changing nature of technology, a universal definition for technology-based items has not been agreed upon in literature (Bryant, 2017). As new item types are being developed all the time, it is difficult to create a definition that takes into consideration an ever expanding range of possibilities. Parshall et al. (2010) used a broad definition which categorised technology-based items as those that employ ‘technologies that use features and functions of a computer to deliver assessments that do things not easily done in traditional paper-and-pencil format’ (p. 215). Similarly, Bryant (2017) asserted that technology-based items refer to items that make use of stimuli, formats and/or response actions *not* associated with traditional, text-based multiple choice and constructed response questions. Applying Bennett’s (2015) conception of ‘first-generation assessments’, standard, text-based MCQs that are delivered using online delivery systems do not constitute technology-based items. It is more appropriate to classify these as *technology-administered items*. In contrast, items such as those found in second or third generation assessments that include, in any way, multimedia objects and/or require response options beyond the selection or insertion of alphanumeric responses are categorised as *technology-based items*. As there is no way to transfer these items to another medium (e.g. paper) they are entirely technology based and should therefore be classified as *technology-based items*.

2.4.1 Classifying Technology-Based Items

The majority of literature on TBAs often erroneously uses the term ‘technology-enhanced items’ or the acronym ‘TEIs’ to describe *all* technology-based items (e.g. Masters & Gushta, 2018; Bryant, 2017). However, this is misleading as their technological features do not automatically ‘enhance’ assessment by expanding or improving construct representation. They merely have the *potential* to do this through their provision of ‘alternative response actions, formatting, types of stimulus and measurement data’ (Bryant, 2017, p. 3). Two components of a technology-based item facilitate construct

representation: the *stimulus piece* and the *interaction space* (Haladyna & Rodriguez, 2013; Measure Progress/ETS Collaborative, 2012; Russell & Moncaleano, 2019). The stimulus piece presents the item's prompt and can include a range of multimedia objects including images, sound, animations or even simulations. The item's interaction space is where the test-taker's actions and responses are recorded. Russell (2016) noted that there are a wide variety of interaction spaces for technology-based items such as 'different types of selected response, drag-and-drop, line and object production, text selection, re-ordering... free-hand drawing, and even the upload of sound, image and video files' (p. 22).

In an attempt to provide some clarity around the types of items that would be used in their newly designed TBAs for US elementary, middle and high-school students, Measure Progress/ETS Collaborative (2012) published a short paper offering some brief guidelines and definitions on technology-based items. Since its release, other academics have proposed similar definitions. Russell (2016) and, more recently, Russell and Moncaleano (2019), has supported the Measure Progress/ETS Collaborative's (2012) approach to separate technology-based items into two categories. The first, *technology-enabled* items, are defined as 'computer delivered items that use digital media as the stimulus (sound, video, or interactive widget), but do not require specialized interactions for response' (Measure Progress/ETS Collaborative, 2012, p. 9). These item types allow for a 'non-traditional' layout of items involving multimedia objects that also use constructed-response and/or selected-response options. This description aligns well with Bennett's (2015) depiction of second-generation TBAs whose items use less traditional item formats '...involving multimedia stimuli, short constructed response, static performance tasks like essays and ... make initial attempts to measure new constructs, beginning to change what is assessed' (p. 371). Russell and Moncaleano (2019) argued that these items are more focused on the stimulus component of the item.

The second category refers to *technology-enhanced* items. These 'include specialised interactions for response and/or accompanying response data' (Measure Progress/ ETS Collaborative, 2012, p. 9). Items within this category are focused on the methods 'used to produce a response to an item... and includes [sic] response actions that differ from selecting from a set of options or entering alphanumeric content' (Russell & Moncaleano, 2019, p. 2). Such response actions include drag-and-drop items, select-text items and shade area. These items are more advanced as they attempt to 'replicate

important features of real environments, allow more natural interaction with computers, and assess new skills in more sophisticated ways' (Bennett, 2015, p. 372). Technology-*Enhanced* items aim to expand construct representation. Bennett's (2015) descriptions of third-generation assessments seem synonymous with the types of technology-enhanced items that testing companies are currently trying to develop.

Figure 2.1 summarises this enabled/enhanced classification system. This classification system is not fully accepted or agreed upon within the field.

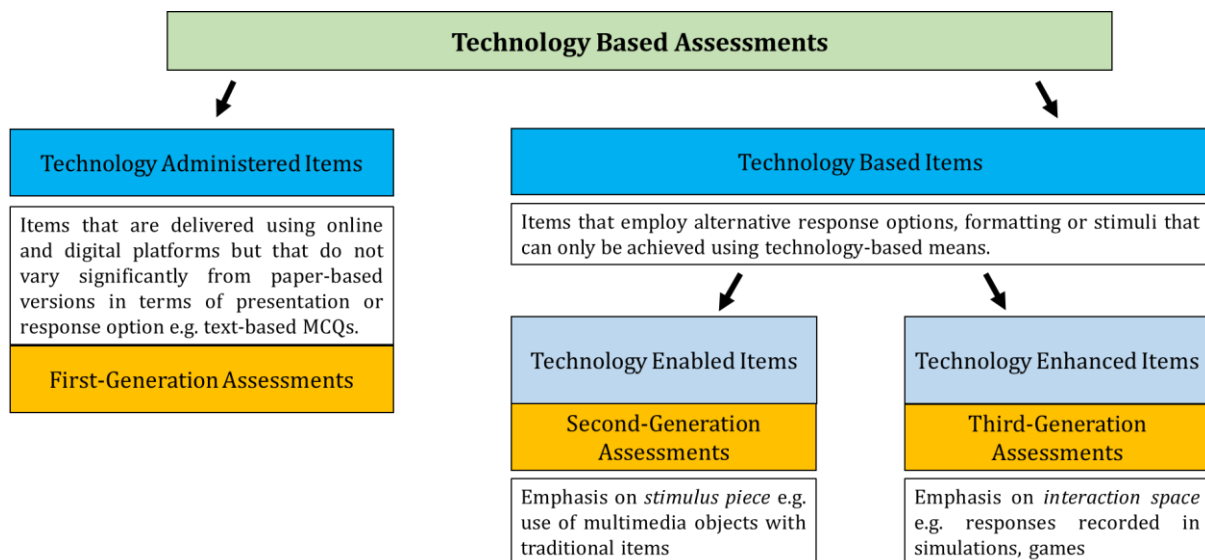


Figure 2.1 A possible classification system for technology-based items (e.g. Bennett, 2015; Russell & Moncaleano, 2019)

While the definitions of technology-*enabled* and technology-*enhanced* items allow for a clear discrimination between the two, this classification system is not widely agreed upon in literature with Russell (2016) noting that a technology-based item can fit into both categories. For example, a recent study by Russell and Moncaleano (2019) analysing technology-*enhanced* items in educational TBAs applied a somewhat restrictive definition of the same. Items involving traditional response options (selected-response, constructed response) were immediately removed from analysis despite including complex multimedia objects or simulations that enhanced construct representation (Russell & Moncaleano, 2019, p. 10). While the authors considered these to be technology-*enabled* (as the technological innovation was limited to the stimulus), it could also be argued that they were technology-*enhanced* items as they ‘replicated [sic]

important features' in the item stimuli (Bennett, 2015, p. 372). This 'blurring' of the lines between what are technology-enabled and technology-enhanced items is becoming more commonplace. This is because some items in TBAs are employing animations (thus making them technology-enabled items) while simultaneously using interaction spaces that are consistent with definitions for technology-enhanced items e.g. simulations. For example, one testing platform commonly used in the credentialing industry, *Surpass* (BTL, 2018), allows test developers to write items with a variety of interaction spaces, as demonstrated in Figure 2.2. These interaction spaces can be accompanied by a range of multimedia objects including videos, animations and interactive images. For some items however, the interaction space and stimulus piece cannot be separated as test-taker responses are recorded in both areas.



Figure 2.2 Interaction Spaces available on the *Surpass* platform (BTL, 2018)

Clear operational definitions that would assist in the classification of technology-based items are unlikely to be available until a better understanding of the range of interaction spaces and their possible relation to an item's stimulus piece is achieved. At present, it appears that classifying items as technology-enabled or technology-enhanced is somewhat problematic. Given the range of media-based stimuli pieces that can be used to create technology-based items, it is unsurprising that there is no formal organisation system using this aspect of an item as a criterion. Reflecting this difficulty, classification systems that focus on the response actions in the interaction spaces of technology-based items have endured in the field. One broadly-cited classification scheme for technology-

based items is Scalise and Gifford's (2006) constraint taxonomy. This taxonomy distils all technology-based items into seven possible categories where the response actions possible within an item's interaction space are arranged on a continuum from select-only to fully constructed responses. In this thesis, an adapted version of Wan and Henly's (2012) categorisation strategy, which is a simplified version of Scalise and Gifford's (2006) constraint taxonomy, will be used. That is, technology-based items will be classified according to the degree of constraint associated with the response actions permissible in an item's interaction space and will include: Selected Response (SR) Items, Figural Response items (FR) and Constructed Response items (CR). Research on each item type will be examined, particularly in relation to their use of static and dynamic item stimuli. Interactive multimedia objects, simulations, can fall into each category, depending on the action required by the test-taker. However, these will be discussed separately given the limited research on their use on assessment contexts.

2.4.2 Selected Response (SR) Items

Traditional selected response (SR) items include a text prompt as a stimulus followed by a range of possible responses from which the test-takers must select the best choice(s). SR items can vary in type and complexity but are characterised by the test-taker *selecting* a response option. SR items include true/false questions, extended matching questions, multiple choice questions and situational judgement items. They are most commonly found in first-generation TBAs as they can be quickly converted from pencil-and-paper items to technology-administered items. Literature in the area asserts that items involving a selection based response action have many advantages including 'efficient administration, automated scoring, broad content coverage and high reliability' (Wan & Henly, 2012 p. 59). Yet, Resnick and Resnick (1992) assert that SR items, and in particular multiple-choice items, tend to decontextualize learning and cannot create an authentic context for the measurement of certain complex skills and competencies, thus limiting their ability to accurately represent a construct.

In response to such criticisms, SR items now include multimedia materials (e.g. images, videos), moving them beyond technology-administered items. Dancy and Beichner (2006) were one of the first researchers that explored the use of multimedia materials in SR items. The authors compared a static and animated version of the *Force*

Concept Inventory Test, a 30-item conceptual test that explores test-takers' understanding of motion and forces using multiple-choice questions. In this study, a parallel version of the paper-and-pencil test was developed by replacing static pictures and descriptions with animated versions. An experimental design where participants ($n = 53$ high school students; $n = 325$ university students) were randomly assigned to complete the static or animated version of the test was used to investigate the impact of stimulus types on test-taker responses. Dancy and Beichner (2006) found that there were statistically significant performance differences between the two conditions in six of the administered 30 items. Those who took the animated version of the test performed significantly better on three of the items than those who completed the static version. An identical finding was found for those who took the static version when compared with the animated version of the test.

In an attempt to explain their findings, the authors hypothesised that the three items that test-takers performed better on when an animated version was available contained some information that could not be communicated by static images or by text alone. Further investigation by the authors found that students with higher verbal ability tended to perform better in the static version of the test ($r = .22$), thus suggesting that the reading ability of test-takers could act as a source of construct-irrelevant variance in testing situations. As a result of their research, Dancy and Beichner (2006) asserted that the animations improved the precision of the assessment. Unfortunately, this is a simplistic interpretation of their findings. The authors themselves noted that 'animations sometimes led students to the correct answer and other times did not' (Dancy & Beichner, 2006, p. 4). When completing test items, test-takers should give an answer that is reflective of their *proficiency* on the construct being measured. In this study, it is unclear if the inclusion of animations improved test-takers' understanding of the question being asked (because it 'removed' the load of having to read and understand text) or if it created a situation whereby the animation disguised any gaps in test-takers' understanding because it provided additional information. Instead, it is more likely that this research suggests that, for SR items at least, animations can alter the outcome of an assessment and that a deeper investigation is needed to understand why this occurs.

Instead of trying to understand the exact conditions where multimedia objects can affect test scores and *why*, recent research is currently attempting to leverage the use of high-definition images and animations to create 'better' SR items. Wright and Reeves

(2016) developed an objective and accurate assessment tool to provide a measure of image interpretation accuracy for radiographers and radiologists. Real, high-quality, high-resolution, musculoskeletal images are presented in the *RadBench* interface to examinees who must decide, using a five-point scale (which practising professionals use), the image's level of normality and if a fracture is present. Images can be maximised and 'zoomed in' on by the test-taker. In a preliminary study involving 42 radiographers and radiologists, the *RadBench* tool was found to be capable of providing benchmark measures of their image interpretation accuracy. The software allows the test-taker to compare their score with the highest, lowest and mean score of others who had engaged with the same bank of items. The authors claim that this approach can be used to provide guidance on the certification status of radiologists and radiographers.

RadBench represents a good example of how the inclusion of images can provide a more realistic context for test-takers to demonstrate specific skills and proficiencies using select-only response actions. This is also evidenced by the use of multimedia Situational Judgement Tests (SJTs). A seminal experiment conducted by Lievens and Sackett (2006) examined the differences between video-based and text-based SJTs that aimed to measure interpersonal skills amongst medical professionals in terms of the test's criterion-related validity. The results of this research study indicated that the video-based SJT was a much stronger predictor of students' performance in medical courses focusing on interpersonal skills ($r=0.35$) than the text-based version of the test ($r=0.09$). These results align well with Sireci and Zenisky's (2006) discussions on construct representation. The use of video as a stimulus in these items broadened the representation of the targeted construct of interpersonal skills. The videos may have provided highly relevant cues such as facial expressions or body language. This may have enhanced the predictive validity to the TBA in question

While multimedia objects can create more contextually appropriate item stimuli for SJT type items, thus increasing their level of construct representation, the interaction space of SR items must also be considered. Efforts by MacCann et al. (2016) to further enhance the construct representation of SR items have recently been made. Using a well-established measure of emotional management, the authors developed a multimedia emotion management SJT that, instead of written descriptions, used acted videos to present both the scenarios and the response options. These response options were hypothesised to have a higher level of construct representation than text-only options as

they better represented the real-world (Figure 2.3). The researchers hoped that the video-based SJT would be a better measure of participants' emotional intelligence. Contrary to this hypothesis, no statistically significant difference was found between the text-based and video-based assessment tools used with the 427 US college students despite the video-based SJT using response options that created a 'real world' context. Both forms of the assessment measured the construct to the same degree. An explanation for this can be found in older research conducted by Haladyna (1999). Haladyna (1999) indicated that unless the 'select-one' approach is present in the 'real-life' application of the skill that is being assessed, as demonstrated in Wright and Reeves' (2013) use of the *RadBench* software, items with selection-based response actions will always be somewhat limited in their capacity to fully represent certain constructs regardless of their use of certain multimedia objects.

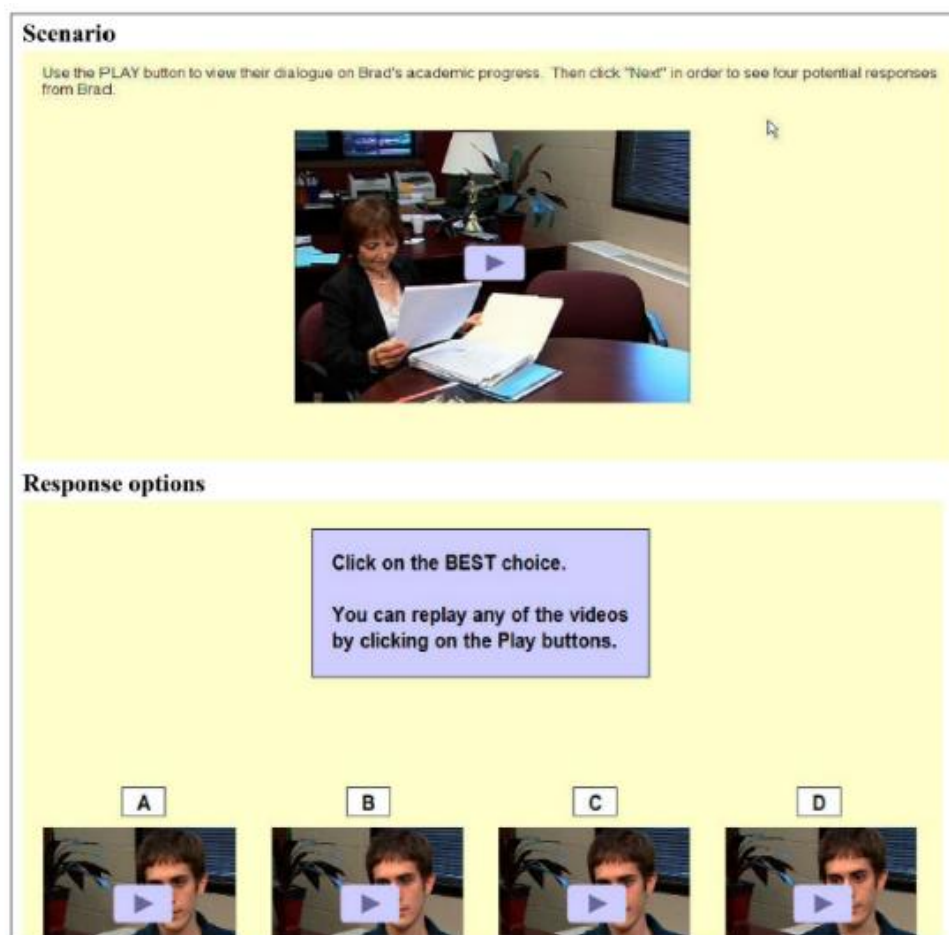


Figure 2.3 Multimedia SJT and response options (Reprinted from MacCann et al., 2016)

2.4.3 Figural Response (FR) Items

Figural response (FR) items require examinees 'to manipulate the graphic elements of an item, click on one or multiple 'hotspots' on an illustration, or complete a diagram by dragging and dropping elements' (Wan & Henly, 2012, p. 63). These items allow examinees 'to rotate, resize, and zoom in or out of a scaled image', select a part of a graphic or drag icons to complete or create a meaningful image (Parshall & Harmes, 2008, p. 9). As a result, FR items are highly dependent on multimedia material such as illustrations, graphs and diagrams. Therefore, there is a greater level of overlap between the item stimulus and the interaction space. As with SR items, FR questions require the test-taker to 'select' their answer. However, FR items are distinguished by the increased level of freedom a test-taker has in what they *do* with the item selected. An early study by Martinez (1991) determined that FR type items in a paper-based assessment of science for 4th, 8th and 12th grade students were found to be more discriminating and more likely to produce more reliable scores than select-only, multiple choice items. Research to support the psychometric value of FR items (such as those in Figure 2.4) in the context of TBAs is still forthcoming.

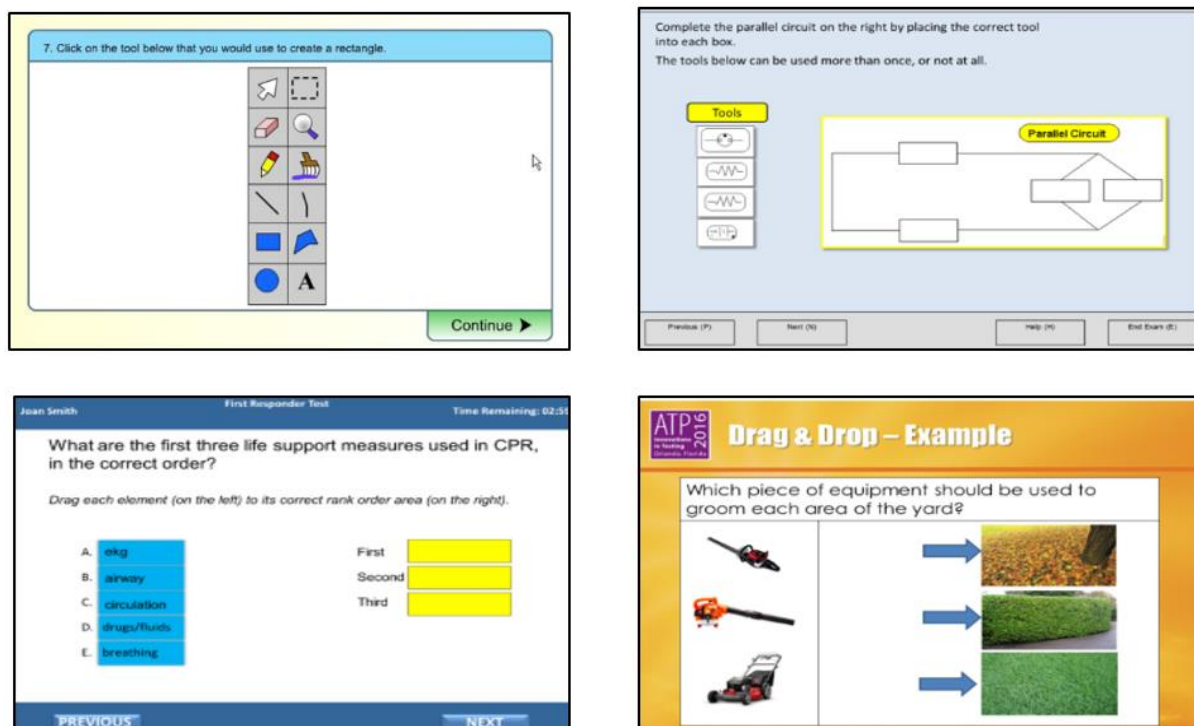


Figure 2.4 Examples of Figural Response (FR) items (Professional Testing, 2018)

Wan and Henly (2012) studied the reliability and efficiency (in terms of test-taker response time) of different item types in TBAs, including FR items that employed the 'drag and drop' and 'hotspot' functions in a TBA interface for school aged test-takers in a state-wide science based achievement test. These FR items presented a 'realistic representation of classroom experiments and real-world phenomena by employing graphics, audio/video media and animation' (Wan & Henly, 2012, p. 62). Using 3-Parameter Logistic (3-PL) Item Response Theory (IRT) Models, FR items were found to be as good as multiple-choice SR items in providing information about test-taker ability. FR items were also considered to be as reliable and efficient as multiple-choice (SR) items. It should also be noted that FR items are less reliant on test-taker reading ability than SR items and, depending on the test-taking population, may be more appropriate items to use if they are 'as good as' SR items. Wan and Henly (2012) did not explore the impact of multimedia type within FR items, making it unclear if the type of multimedia object used influenced the function of the FR item. Another study by Kong et al. (2018) also found significant differences in response times for different item types. These differences were amplified according to the device type (tablets, computers etc.,) used by test-takers. Students ($n=974$) in the tablet condition of the study took longer to respond than students in the computer condition for hot spot items (3.59s longer).

Another FR item type that is receiving increased attention in medical education are concept maps. Concept maps are graphical representations of knowledge and have been used to promote 'meaningful learning, critical thinking and problem solving skills' (Ho et al., 2018, p. 2). Features of concept maps include a hierarchical structure and cross links (Figure 2.5). These characteristics, as well as recent advancements in web software, have made automated test-scoring of these items possible. An online mapping tool called *Knowledge Maps* (Ho et al., 2018), uses the weighted proposition method, as developed by Chang et al. (2005) where each concept (also described as a 'node') in the teacher's map is given a weight from zero to one. More important 'nodes' are given increasingly higher weightings. The student's score is calculated by comparing their selected or constructed nodes to those in the teacher's map. Correct answers are given a score of one, partially matched answers given a score of 0.5, and missing answers are awarded zero. In their pilot study involving first-year medical students ($n=137$), Ho et al. (2018) compared students' scores on concept maps such as these to their grades for a modified essay question on the same topic. Students were presented with an incomplete concept

map where dropdown lists of options for each node (or concept) were available. Although the students employed a 'select-only' response action, test-takers could also move nodes if required thus fulfilling the criteria for a response action consistent with the definition for an FR item. Similar gaps in understanding were elicited using both assessment items. Furthermore, 'nodes' were highly consistent in discriminating between students of different abilities. Cronbach's α also demonstrated high internal consistency across the nodes in the map ($\alpha = 0.77$).

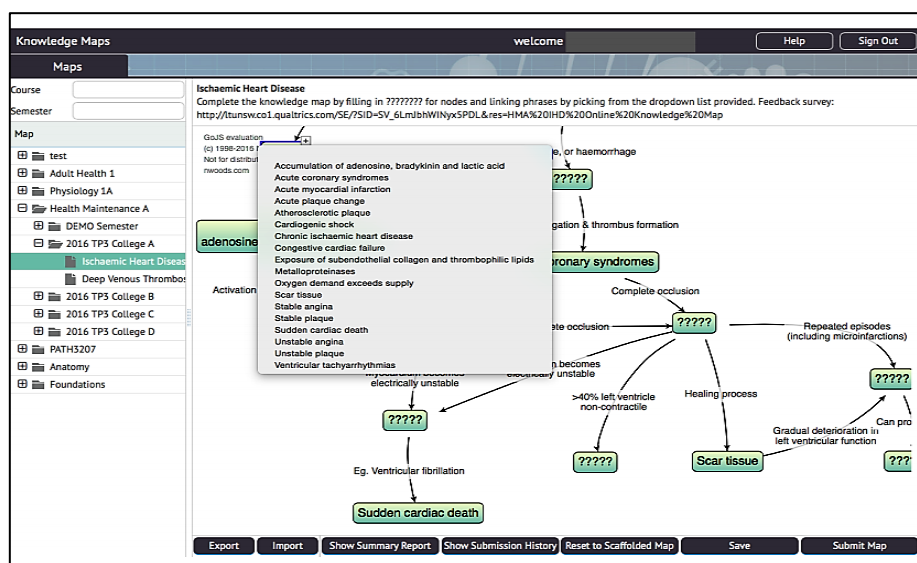


Figure 2.5 Sample interface from *Knowledge Maps* (Reprinted from Ho et al., 2018)

FR items have received limited attention in research. For example, a comparison between FR items that use and do not use multimedia stimuli has not been conducted. This lack of research was initially documented by Wan and Henly (2012) when their review on these item types noted that literature discusses the topic extensively but provides a limited empirical evaluation of FR items. What research currently exists (e.g. Martinez, 1991; Woo et al., 2014) indicates that they are 'as good as' SR items. There is an advantage to FR items however, as FR items appear to provide richer diagnostic information¹⁰ than the SR format. While it is possible to infer from the incorrect option selected in a multiple-choice item what a test-taker's difficulties and misconceptions are, it is easier to record, analyse and assess cognitive processes and problem-solving

¹⁰ Assessments that are constructed on the basis of appropriate models of learning can provide crucial diagnostic information. According to Leighton et al. (2010), diagnostic information from assessments informs educators about students' strengths and needs. This can then support 'meaningful learning' as students and teachers have a deeper understanding of what future instruction and learning should include.

strategies in FR items as the test-taker must ‘do’ something with the response option that has been selected or identify for themselves what must be selected.

FR items can provide information on whether a test-taker knows an answer or not but also the test-taking strategies that they employ. Yet, certain design considerations may influence what information can be extracted about a test-taker’s performance. To understand the effect of the ‘drag-and-drop’ FR item design on test-takers’ strategy use, Arslan et al. (2019) conducted a four-condition experiment with 378 technology professionals. ‘Drag-and-drop’ FR items have two key design features: sources (text/images that can be manipulated and moved) and targets (where the sources must be placed according to a particular rule). The online assessment items used in the study involved different mathematical questions, including calculating the area of a shape. The researchers explored the value of four different design strategies to FR items: a) presenting sources and then targets b) presenting targets and then sources, c) swapping the content of the sources and targets and d) no problem statement; (Figures 2.6a, 2.6b, 2.6c, 2.6d).

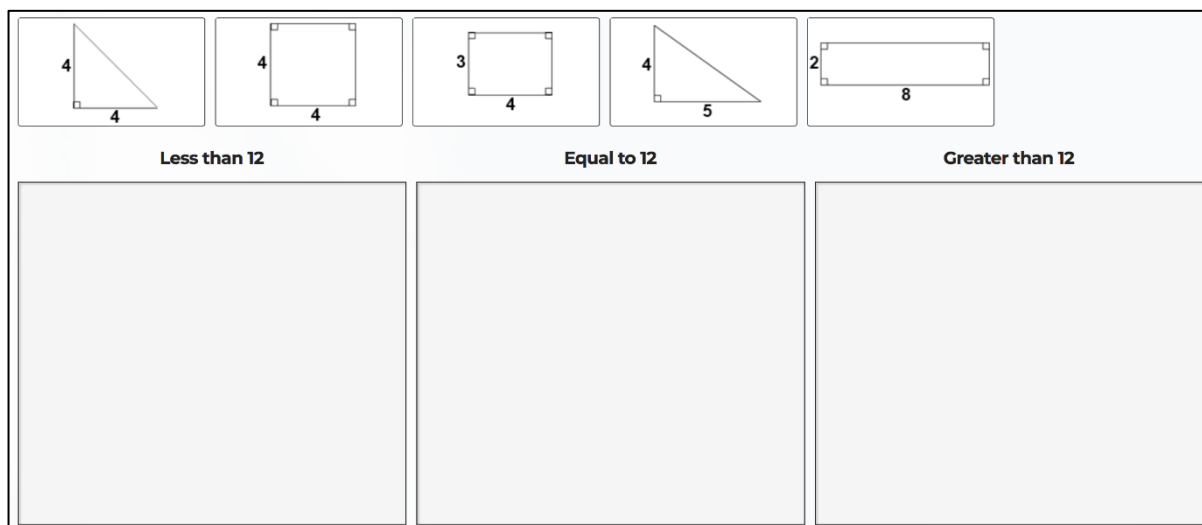


Figure 2.6a Sources first (Reprinted from Arslan et al., 2019)

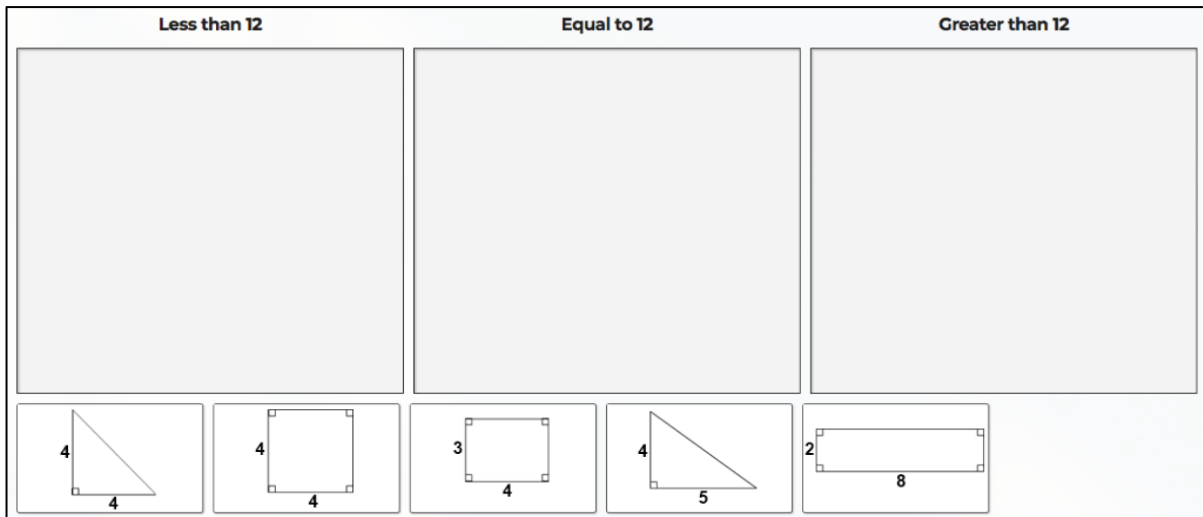


Figure 2.6b Targets first (Reprinted from Arslan et al., 2019)

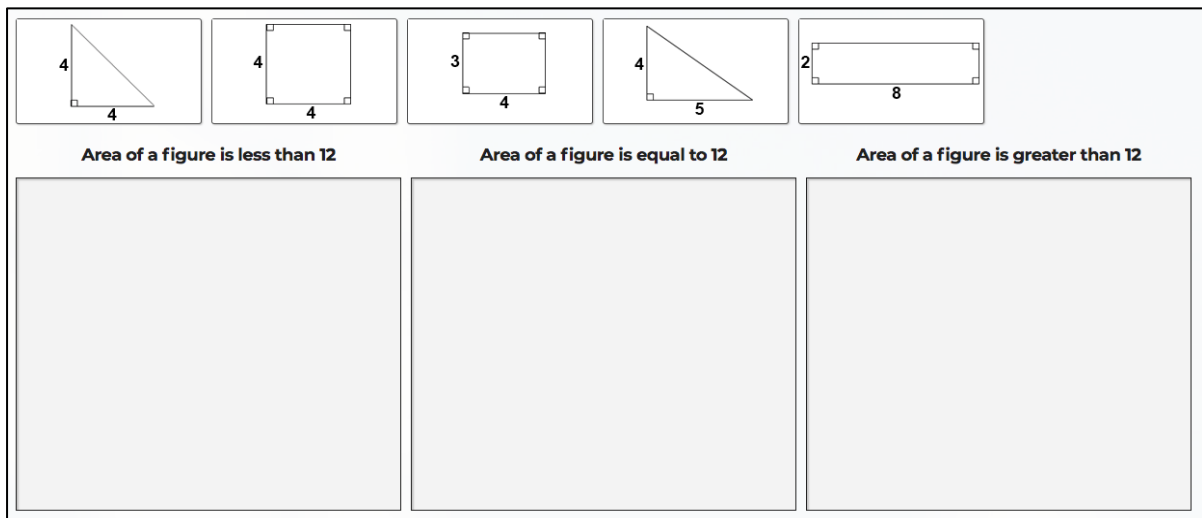


Figure 2.6c Swapped content (Reprinted from Arslan et al., 2019)

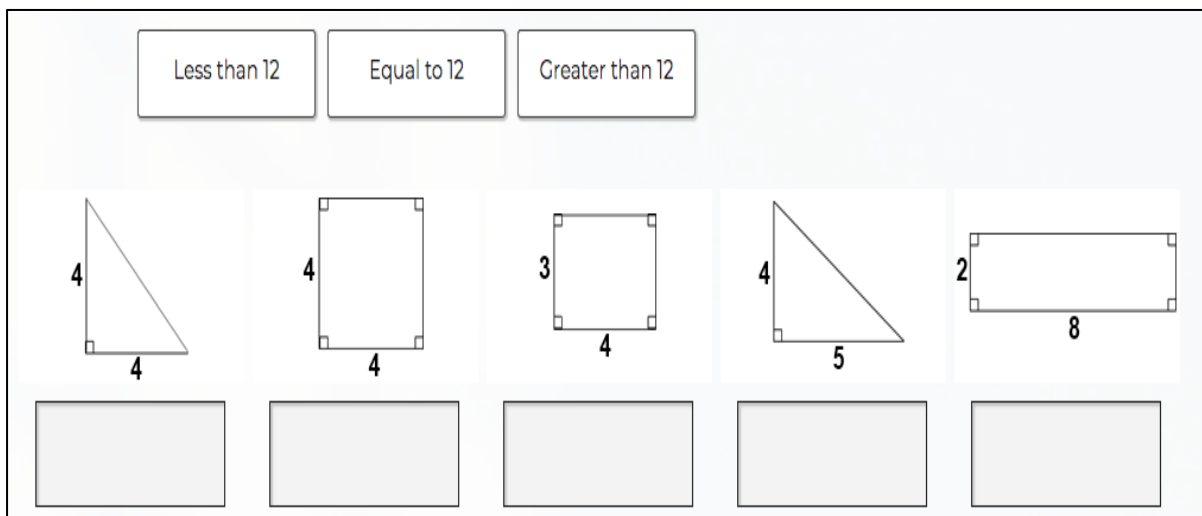


Figure 2.6d No problem statement or instructions (Reprinted from Arslan et al., 2019)

Arslan et al. (2019) found that test-takers' response strategies were affected by the design of the drag-and-drop FR item. For example, in the sources and target first conditions (Figures 2.6a and 2.6b), test-takers were more likely to adopt a *source focus approach* whereby cognition is 'offloaded' by action i.e. test-takers focused on organising the sources according to the target rule (e.g. total area of the being more, equal to or less than 12cm²) rather than seeing the targets as separate items that should be dealt with one-by-one. This is considered to be a more efficient strategy and one that test-takers will engage in if the item is designed according to one of these conditions. In contrast, when a swapped content approach was adopted (Figure 2.6c), whereby the 'rule' was represented as the source to be organised, test-takers were more likely to engage in a target focus approach, possibly because there was a greater amount of physical distance between the sources and targets. This has important implications for the design of FR items as it seems that understanding which information should be presented as a source and which information should be presented as a target needs to be carefully considered at the start of the item writing process. Further research to determine the most efficient design of these items is still required at the time of writing.

2.4.4 Constructed Response (CR) Items

Wan and Henly (2012, p. 63) define constructed response (CR) items as those which require students to create an alphanumeric response which can vary in length. These include short or extended written responses (e.g. fill-in-the-blanks, essays, spreadsheets). Although the CR item type is considered to be very traditional and most commonly associated with paper-based assessments, it is still used extensively in TBAs (BTL, 2018). CR items are accepted as being capable of providing a strong and reliable measure of student knowledge (e.g. Livingston, 2009). This can explain their continued use in TBAs. Their future use in TBAs is also secured as modern technology can be used to improve construct representation in CR items. Test-takers can construct their own responses to an item in a separate application and then upload their work to be assessed. For example, a TBA may aim to measure a test-taker's proficiency in managing spreadsheets in the Excel programme. Allowing test-takers to upload an authentic sample of work can 'fully represent' the desired construct better than SR or FR items as it cannot 'leave something out' (Messick, 1988, p. 34) and is a genuine representation of the

construct. For instance, consider exams in foreign languages. If the skill to be assessed is conversational proficiency in Greek, an SR item that requires the test-taker to match an image to the correct word may assess aspects of this skill (e.g. vocabulary). Yet, it cannot completely address an examinee's ability to hold a conversation in Greek. In contrast, a CR item that incorporates some multimedia functionality, which could include sound recording or audio playback, may provide a greater level of construct representation if the test-taker responds to an audio prompt by recording a response that can later be scored. Unfortunately, little published research comparing the value of CR items involving different multimedia stimuli and/or response options exists to support such a hypothesis.

CR questions are used in assessments as they are thought to measure test-takers' ability to apply, analyse, evaluate, and synthesize their knowledge (Downing & Haladyna, 2006). This has been contested though as some argue that essays and 'fill-in-the-blank' items may not reflect the authentic context of the construct being measured, particularly in relation to skills such as problem solving or collaboration (NCCA, 2007). Consequently, newer technology-based items have emerged to address these concerns; the most significant of these being simulation-type items.

2.4.5 Simulation-Type Items and Tasks

Levy (2012) defined simulation-type items as those in which the test-taker 'is presented with, works with, or produces a work product that contains a simulation of a real-world scenario' (p. 10). Simulations are a type of 'interactive' multimedia object that usually involve test-takers engaging or manipulating text and/or some form of pictorial representation to create the required work product (Levy, 2012). Their use in educational TBAs, particularly for science-based tasks, is growing (Levy, 2012; Greiff et al., 2018). Simulation-type items in these TBAs often require test-takers to engage in actions associated with SR or FR items to produce an output ('work product') from the simulation that is to be scored on its own or in conjunction with another item or series of items (e.g. OECD, 2018; Iseli et al., 2010). In this way, simulation-type items 'blur' the lines between the already problematic *technology-enabled* and *technology-enhanced* item classifications. For these items, it is difficult to differentiate between their interaction spaces (where test-taker actions and responses are recorded) and stimulus pieces (that

which contains the prompt or ‘media’). For some, items that include simulations should not be termed ‘items’ at all and should instead be called ‘tasks’ (Almond et al., 2014) as they ‘yield complex work products that generate *multiple* observed outcomes’ (p. 2). For example, a ‘task’ may include reading a passage of text, executing a range of simulations and then answering a range of ‘traditional’ SR or CR items. For the sake of clarity and consistency, ‘simulation-type items’ will be used throughout the thesis but the term ‘tasks’ may also be used where appropriate.

Quellmalz et al. (2013) developed a series of simulations for 12-year-old students that assessed their knowledge of ecosystems. The authors noted that some technology-administered items, including multiple-choice items, can effectively measure declarative knowledge such as scientific facts or definitions, but that they could not provide evidence of science inquiry practices such as making observations or designing and conducting investigations. Using simulations, Quellmalz et al. (2013) examined the inquiry abilities of over 1500 ($n=1566$) middle school age children across three modalities with different response actions (Condition 1: Static Images, Multiple-Choice items; Condition 2: Animated Videos, Multiple-Choice Items; Condition 3: Interactive Simulation, Designing/Running population trials); Figure 2.7.

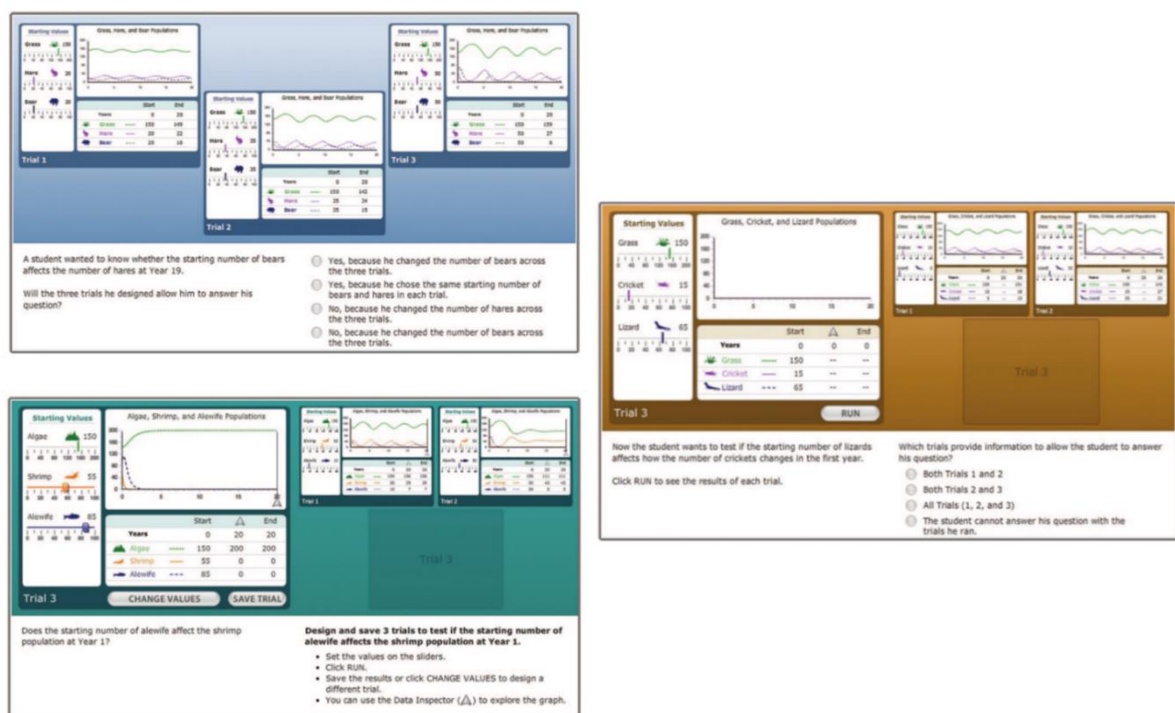


Figure 2.7 Image, animation and simulation conditions (Reprinted from Quellmalz et al., 2013)

Each condition was designed to assess three key science constructs – declarative knowledge (knowing '*what*'), schematic knowledge (knowing '*why*') and procedural/strategic knowledge (knowing '*how to use and apply*'). In this study, a combination of methods – a generalisability (G) study, Multi-dimensional Item Response Theory (MIRT), and confirmatory factor analyses – examined the measurement properties of the chosen modalities and response actions. All three statistical analyses found that the differences among the assessment modalities were relatively small when measuring for the declarative and schematic constructs. However, the authors found that the interactive (simulation) condition was more effective in measuring the procedural/strategic knowledge. The MIRT analysis found a higher reliability coefficient (.82) for the construct related to procedural/strategic knowledge in the interactive modality when compared with the animated (.77) or static conditions (.77). Quellmalz et al. (2013) used these findings to support their claims that simulation-type items better represent the construct of procedural and strategic knowledge.

Quellmalz et al.'s (2013) work provided crucial information on the use of simulations in educational assessments. However, further research on the use of assessment evidence embedded in simulations has been somewhat slow to emerge with some exceptions (e.g. Shute & Ventura, 2013; Shute et al; 2016). Given that Quellmalz et al. (2013) showed how items involving different multimedia stimuli (images, animations, simulations) can change how test-takers interact with items, this is a significant oversight. Even more so when one considers that Quellmalz et al. (2013) showed how multimedia objects can affect construct representation. Therefore, simulation-type items, as they are currently designed, need careful consideration and examination in future research as do all items that include multimedia objects.

2.5 Commentary on Technology Based Items

Figure 2.8 summarises the classification system for technology-based items used in the current thesis, as informed by Russell (2016), Russell and Moncaleano (2019), Wan and Henly (2012), Bennett (2015) and Bryant (2017).



Figure 2.8 Classification System for Technology-Based Items

Regardless of the classification system used, it is important to remember that an item's capacity to measure a construct is dependent on 'how well the item fits with the theoretical representation of that construct' (Messick, 1988, p. 15). As demonstrated by Lievens and Sackett (2006), the inclusion of multimedia objects or greater computer functionality in the prompt or interaction space of an item can allow for a clearer representation of a construct as it would appear in the real world. Although there are some concerns regarding how well the interaction space represents a construct when response actions are limited to select-only or alphanumeric options, as discussed by Russell (2016), the inclusion of multimedia objects in these items appears to positively

contribute to construct representation. This increased level of construct representation can also be accompanied by an increased level of face validity.

Face validity refers to the extent to which a test appears to measure what it is intended to measure (Ward et al., 2013). If there is a consensus among stakeholders other than the test-developer that the test's items seem to measure what the test proposes to measure, the test has strong face validity. Ward et al. (2013) claimed that tests with strong face validity can encourage people to respond to or have an improved perception about the testing experience. Kanning et al. (2006) created a video-based SJT where even the response options of the items were presented with the use of acted videos. The purpose of their study, conducted with 284 police officers in Germany, was to examine whether such an approach would elicit more favourable test-taker experiences. After administering a selection of SJT items in three different formats (a text-based SJT, an SJT with video prompts and text-based response options, an SJT with video-based prompts and response options), the researchers measured test-takers' attitudes towards each test. The text-based items were perceived to have the lowest levels of job-relatedness, an indicator of low face validity, and fairness.

Despite these apparent improvements in face validity, there are some risks associated with the inclusion of multimedia objects in technology-based items. The inclusion of multimedia objects can provide a more realistic context for test-takers to demonstrate specific skills and proficiencies but they may undermine the construct being assessed. Malone and Brünken (2013) found an interaction effect between multimedia type (animated images, static images) and expertise level in an experiment that compared the performance of novice and expert drivers ($n=100$) in a TBA involving multimedia (instead of traditional text-based vignettes) stimuli for multiple choice items. Novices benefitted from the animated presentation as they did not have to infer relationships and motion from a static image. Yet, the animations did not allow for the clear identification of expert drivers. These drivers outperformed novice drivers in the static image condition only. This provides some evidence for the argument that the inclusion of multimedia objects in technology-based items may influence test-taker behaviour and performance. This could then modify the evidentiary structure of an item's key measures (e.g. discrimination, difficulty etc.,) and the types of inferences that can be made from them.

Simulations as interactive multimedia objects may further interfere with the assessment process if their interfaces or engagement mechanics are irrelevant or

interfere with the assessment of a construct or criterion (Rupp et al., 2010). Objects in the environment, the interface, or colours that increase engagement or create a sense of authenticity, may also introduce construct-irrelevant variance that can negatively impact the sensitivity and specificity of the item as an assessment tool. Ironically, an item with a high level of construct fidelity could interfere with construct representation (and thus, validity) if potential construct-irrelevant variance created by poor interface design or other implementation challenges are not taken into consideration. For example, to demonstrate their word processing skills, a student may be asked to type up a short paragraph. Russell (2016) argued that this test item has a high level of construct fidelity. However, if the word-processor used by the testing environment has unfamiliar icons, the design of the interface itself may interfere with test-taker engagement, despite the high level of construct fidelity. Russell (2016) noted that a commitment to construct fidelity in the construction of test-items needs to be considered in conjunction with a number of other factors. To determine if a technology-based item, is worth the cost and effort of development, Russell (2016) designed the TEI Utility Framework.

2.5.1 TEI Utility Framework (Russell, 2016)

According to Russell (2016), a technology-based item's utility¹¹ can be quantified to signify how well a given interaction can collect evidence about a particular construct in an accurate, efficient and high-fidelity manner. If an item has a high level of utility, it is worth the cost and effort of development. Russell (2016) designed this TEI Utility Framework to help identify an item's value for test developers. Three characteristics influence the utility of a technology-based item with this framework: a) construct fidelity, b) usability and c) accessibility. *Construct fidelity* is the 'product of the context created through the interaction, the interactions itself and the targeted construct' (Russell, 2016, p. 24). Simulations can have a high level of construct fidelity as they allow the test-taker to demonstrate response actions that are reflective (or at least representative) of the construct in the real-world environment e.g. selecting variables. Therefore, a high level of construct fidelity is observed when an item resembles the real-world scenario that the

¹¹ In many ways, the term 'utility' in Russell's (2016) framework relates very closely to the concept of validity. However, there are some key differences between the two. Validity refers to the degree to which test evidence supports the interpretation of test scores (Messick, 1994). In contrast, utility attempts to assign value to individual test items using a triarchic approach that includes validity (e.g. construct fidelity), usability and accessibility. While validity is considered an important aspect of this framework, it is not the only factor considered when deciding the value of a test item. In contrast, validity is only concerned with how well an item can gather evidence to support score interpretations.

construct is associated with and allows the test-taker to act and behave in a way that produces an authentic response that represented the targeted construct. *Usability* considers how well a TBA can allow test-takers to efficiently produce responses i.e. ‘How easily can a novice user produce the desired response?’ The final part of this framework, *accessibility*, refers to how well a test-delivery system allows test-takers with specific needs (e.g. motor skills, vision impairments etc.,) to produce responses.

Different levels of each of these three components interact to create a utility level for an item. For example, if fidelity, usability and accessibility are low, the interaction space has low utility. It is poorly aligned with the construct being measured, implementation is inefficient and the item is difficult for some test takers to engage with. It is not worth the cost of development. When there is a discrepancy between the three components, such as when usability and accessibility are high but construct fidelity is low, then the item may still have a moderate level of utility. Unfortunately, the ‘directness of the inference’ (Russell, 2016, p. 29) is somewhat reduced. Russell (2016) provides an example to explain this. A test-taker may be asked to ‘drag-and-drop’ sentences into an order that reflects the plot of a given story. The response actions involve the test-taker to be able to select, drag, and position content. This creates an interaction space that is unrelated to reading comprehension in the ‘real-world’. Reordering sentences is not a context in which test-takers usually apply their understanding of a text. The construct fidelity of this item is therefore low. Yet, the evidence provided by the ordering of sentences supports an inference about the test-takers’ comprehension of the events of the story. The interaction space is easy for the test-taker to use and can be implemented in an efficient and accessible manner. As a result, the utility of that interaction for measuring reading comprehension can be deemed moderate or adequate. Similarly, if the items level of construct fidelity is high but the item’s usability and accessibility is low, the item’s overall utility is low. Although the skills elicited in the item are closely associated with the target construct, implementing the item poses significant challenges for test-takers in terms of usability and accessibility. Therefore, the overall utility of the item is low and should be carefully considered before it is invested in.

Russell’s (2016) TEI framework is most suited to inform ‘value for money’ decisions during the test development process. Yet, its emphasis on construct fidelity illustrates how the value of a technology-based item is closely related to how well that item can represent a construct. Construct representation is dependent on what is

included in the item's stimuli and what is required of the test-taker in the interaction space (Russell, 2016; Russell & Moncaleano, 2019). Given the increased use of multimedia objects in the stimuli of test items, understanding how multimedia objects affect test-takers' performance and behaviour, and indeed the construct being examined, seems particularly relevant at this moment in time. The overview of research presented here suggests that technology-based items with complex multimedia features have been incorporated into assessments without a clear appreciation of their 'differences, measurement implications, cost-benefit trade-offs, or effects on test-takers' (Bryant, 2017, p. 1). A better understanding of the impact of multimedia objects on an individual's cognitive processes is required to inform decisions regarding the role multimedia objects can and should play in technology-based items. Therefore, a cognitive theory of multimedia assessment needs to be explicated.

2.6 Towards a Cognitive Theory of Multimedia Assessment

Various types of static and dynamic pictures can be found in digital assessments as complex information can be easily displayed in this format to test-takers through the use of images, animations and videos. The use of pictorial information in assessment situations is unsurprising given the widespread use of instructional videos and audio-visual presentations for *learning*. This trend in educational instruction emerged as a result of research which supports the presence of the *multimedia principle of learning*, where learner performance improves when learning occurs with text and pictures rather than text alone (Mayer, 2017; Cohen's $d = 1.67$)¹². Yet, Kirschner et al. (2016, p. 1) have queried if this multimedia principle can be applied in testing and assessment contexts. Similarly, Lindner et al. (2017a) argued that in order to use pictorial elements to their full potential when devising *assessments*, research on the behaviours and performance of *test-takers*, rather than learners, must be undertaken¹³. Researchers such as Lindner et al. (2017a; 2017b) investigated if theories of multimedia learning can be applied to the

¹² 'Words' refers to verbal forms of information, which include printed or spoken text. 'Pictures', according to Jamet et al. (2008), can encompass a variety of possibilities, including, amongst others, static and animated illustrations, graphs and diagrams.

¹³ The majority of research in education has been based in instructional rather than assessment contexts. Therefore, it is possible that some findings may not be generalisable between the two. However, it is also important to acknowledge that the fields of assessment and learning in education cannot be rigidly separated – effective instruction requires good assessment practices to inform learning (NCCA, 2007). Research within both settings is required in order to ensure that the process of assessment and learning complement each other.

context of testing. The Cognitive Theory of Multimedia Learning (CTML; Mayer, 2009) is the leading theory in this field and is the basis of much research conducted to date. To determine if this particular theory is an appropriate one to inform current efforts to develop a cognitive theory of multimedia assessment, an overview of the CTML is required at this juncture.

2.6.1 Cognitive Theory of Multimedia Learning (CTML)

Mayer (2009) asserted that the occurrence of the multimedia principle can be explained by applying what is already known from decades of research from the field of cognitive psychology about how people acquire information. With this in mind, Mayer (2009) constructed a cognitive model of multimedia learning that represents how the human information-processing system works when presented with multimedia materials. Figure 2.9 (adapted¹⁴ from Mayer, 2009, p. 61) illustrates this model. The model depicts memory stores; specifically, sensory memory, working memory, and long-term memory. Pictures and words come in from the outside world as a multimedia presentation (A) and enter sensory memory (B) through the eyes and ears. This ‘sensory memory allows for pictures and printed text to be held as exact visual images for a very brief period in a visual sensory memory ... and for spoken words ... to be held as exact auditory images for a very brief period in an auditory sensory memory’ (Mayer, 2009, p. 62). In this way, there are two processing channels: auditory/verbal (yellow channel) and visual/pictorial channel (pale blue channel). Mayer (2009) posited that on-screen or printed text passes through both channels as it is verbal information perceived by ‘seeing’ the word or image represented by a word and by ‘hearing’ it in their head.

¹⁴ Figure 2.9 represents a simplified version of Mayer’s (2009) original explanation of the proposed cognitive model of multimedia learning. Additional markers have also been included in Figure 2.9 e.g. A, B etc., to guide readers.

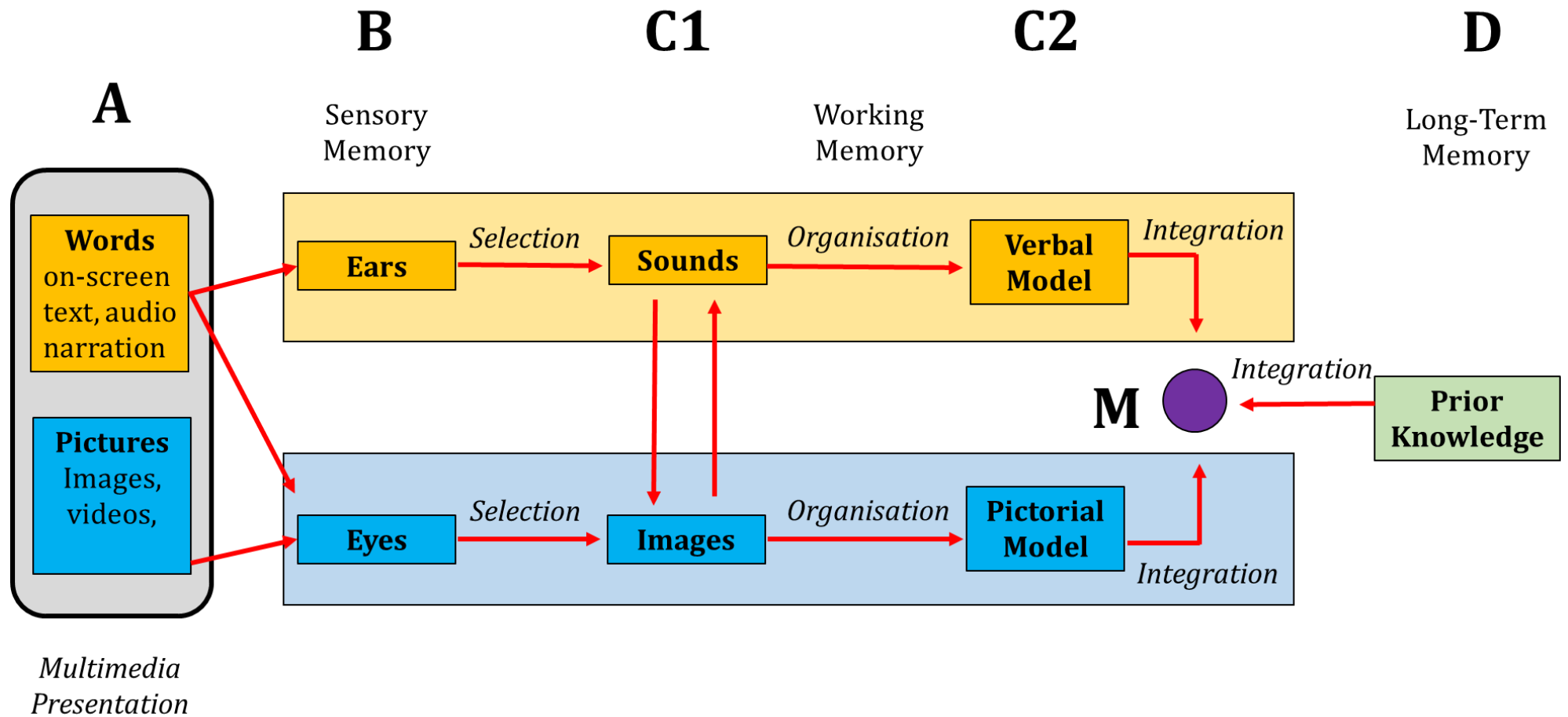


Figure 2.9 Cognitive Theory of Multimedia Learning (adapted from Mayer, 2009, p. 61)

Cognitive psychology posits that working memory is used for temporarily holding and manipulating knowledge in active consciousness (Sternberg, 2009). In Figure 2.9, the initial section of working memory (C1) selects the relevant raw material that comes into working memory based on the two sensory modalities (ears and eyes). Mayer (2009) explained that the arrow from sounds to images (C1) represents the mental conversion of a sound (such as the spoken word “rat”) into a visual image (such as an image of a rat). Similarly, the arrow from images to sound represents the mental conversion of a visual image (such as the printed word “cat” or a mental image of a cat) into a sound (such as the sound of the word “cat”) – that is learners mentally hear the word “cat” when they see a picture of one or read the word. In this way, Mayer (2009) used a ‘representation mode’ approach to illustrate how verbally and pictorially based models in working memory can be constructed from information that was processed in different channels. The latter stages of working memory (C2) represents the knowledge constructed by the learner in working memory, specifically the verbal and pictorial mental models and the links between them to create a single representation or model (M) of the information contained in the multimedia presentation. This model is also informed by the learner’s prior knowledge of the topic as contained in their long-term memory (D).

Three main assumptions based on cognitive psychology research underpin this explanation of the multimedia principle which is referred to as the Cognitive Theory of Multimedia Learning (CTML; Mayer, 2009): the dual-channels assumption, the limited capacity assumption and the active processing assumption. These are summarised in Table 2.1 and are each explained in more detail in Appendix A.

Table 2.1 Three assumptions of the Cognitive Theory of Multimedia Learning (CTML)

Assumption	Description
Dual Channels	Visual and auditory information are processed separately.
Limited Capacity	There is a limit to the amount of information that can be processed in each channel at one time (cognitive load).
Active Processing	Learners must <i>select</i> the relevant information, <i>organise</i> the information into a coherent model in order to <i>integrate</i> with previously acquired knowledge.

The three main assumptions contained in Table 2.1 that underlie the CTML have been extensively researched within the area of cognitive psychology as they allow for empirically tested hypotheses to be devised. This is evidenced by the range of research available as summarised by Butcher (2014) and Clark and Mayer (2016). As a result, ten evidence-based design principles for the creation of technology based multimedia materials – which use these three assumptions as a basic guide – have also been created and researched. These principles are consistent with the propositions contained within the dual channels, limited capacity and active processing assumptions. These principles have also been devised based on the results of several different experimental studies (Mayer, 2017). They are outlined in Table 2.2.

Table 2.2 Ten design principles for multimedia methods of instruction in computerised learning environments (Mayer, 2017)

Principle	Explanation
Coherence	Irrelevant information should not be included.
Signalling	Signals should be used to help guide the learner to relevant information.
Redundancy	Audio narrations should not be accompanied by on screen text.
Spatial Contiguity	Images and labels should be presented side by side.
Temporal Contiguity	Audio narrations should be synced with each segment or event in a multimedia resource.
Segmenting	Information should be presented in short ‘chunks’ or segments.
Pre-training	Key words should be presented to learners before viewing multimedia materials.
Modality	Text based information should be conveyed aurally.
Multimedia	Words and pictures should be presented together.
Personalisation	Informal language should be used.

2.6.2 Applying the CTML to assessment

When computer-based learning materials involving multimedia objects were in their infancy, researchers immediately realised the implications of this development and the necessity of a coherent approach to research to inform practice. This was achieved through the CTML (Mayer, 2009). It is logical to assume that if multimedia *learning* involved more than the simple transference of paper-based learning principles to a computer screen, so too does multimedia *assessment*. Yet, the creation of a cognitive theory directly related to multimedia assessment is still in its infancy with Kirschner et al. (2016) being the only real proponents of its development. Instead, many researchers are applying the CTML to testing and assessment contexts. Lindner et al. (2017a), for example, argued that ‘both learning and testing require students to encode and understand the given information... to build a coherent mental model’ (p. 483). Therefore, applying the assumptions and principles of the CTML to research involving multimedia objects in assessment scenarios may, at first glance, appear appropriate. Yet, it is incorrect to assume that all aspects of the CTML can be transferred to an assessment-based context. In fact, the application of the CTML and its principles to TBAs involving multimedia objects could prove problematic, particularly in relation to the role of cognitive load, construct measurement and the expertise-reversal effect.

2.6.2.1 The Role of Cognitive Load

Mayer’s CTML (2008; 2014) is heavily influenced by the assumption that cognitive systems involved in the processing of information are limited in their capacity¹⁵. This suggests that learning can be negatively impacted when cognitive ‘overload’ occurs and working memory capacity is exceeded. As a result, the CTML recommends managing the cognitive load that learners experience when engaging in multimedia learning. However, minimising all possible sources of cognitive load in assessment materials would reduce a test item’s complexity or discriminatory ability. In the case of learning materials, the emphasis has traditionally been on minimising cognitive load. In contrast, it may be more appropriate to seek a threshold level of cognitive load to ensure a desired level of difficulty or discrimination in test items. An item should have an ‘optimal’ level of cognitive load so that it can be used to discriminate between test-takers. As noted by Kirschner et al.

¹⁵ This assumption is draws heavily on Sweller’s (Chandler & Sweller, 1991) Cognitive Load Theory.

(2016), 'a constructive dilemma exists between fostering instructional understanding by reducing extraneous load and ensuring ecological validity in assessment by keeping this load relatively high' (p. 19). For example, the ten design principles (Table 2.2) of the CTML aim to maximise learning by helping learners to minimise cognitive load. One design principle from the CTML that aims to achieve this is the signalling principle, where relevant material is highlighted to the learner using visual cues e.g. arrows. In line with the limited capacity assumption, this facilitates a reduction in cognitive load as the learner does not have to process any unnecessary information. If this principle is applied to the design of multimedia assessment materials, then test-takers should also be made aware of the key points of the item using some form of cueing system.

Kirschner et al. (2016) question the application of Mayer's (2014; 2017) multimedia design principles being applied to assessment and tests, querying if testing materials should be designed to minimise cognitive load. This is because *managing* cognitive load is a key aspect of tests. In learning, the primary purpose of instructional materials is to help learners store and process information (Mayer, 2014). In contrast, tests require test-takers to retrieve the necessary information from memory and then apply their knowledge to complete a task (Kirschner et al., 2016). For example, one of the most basic purposes of a test is to identify competence and to gain an indicator of an individual's overall skill (Haladyna & Rodriguez, 2013). Selecting relevant information in complex tasks can be an indicator of competence. Signalling to a test-taker what information should be attended to, rather than allowing them to select it themselves, may reduce the criterion related validity of the item.

2.6.2.2 Construct Measurement

Other research also indicates that multimedia items in TBAs need to be carefully applied to ensure that construct measurement occurs as intended. In a small-scale, mixed methods study conducted by Vorstenbosch et al. (2014), seventeen first-year medical students answered Extended Matching Questions (EMQs) regarding their understanding of gross anatomy, using either labelled images or answer lists in a paper-and-pencil test. They also orally outlined their strategies and thoughts on answering these items using the 'Think Aloud' research approach. Vorstenbosch et al. (2014) found that EMQs with and without images seemed to 'measure different skills, making them valid for different testing purposes' (p. 107). Students used more cues from EMQs with images and

visualised more often in EMQs with text-based answer lists. Items without images seemed to test the quality of students' mental images while questions with images tested their ability to interpret visual information. These findings suggest that the inclusion of multimedia objects in an item can modify the construct being assessed which may have significant implications for any interpretations or judgements.

More recent research involving TBAs by Lindner et al. (2017a; 2017b) further supports this assertion. In a classroom-based experiment involving 410 students (10-12 years old), Linder et al. (2017a) found that the inclusion of images in items improved all students' performance when compared to items that contained no images (text-only). Using the results from their generalised mixed effects model, the authors termed this significant positive main effect of pictures on student performance the 'multimedia effect in testing'. A cognitive facilitation effect was also noted, whereby items containing images accelerated the item solving process. While Lindner et al. (2017a) claimed that this study indicates that images in TBAs could promote more reliable test scores, and thus support a more valid interpretation of students' achievement levels, further research is required for a number of reasons. For example, in relation to illustrations that can accompany text, Carney and Levin (2002) noted that there can be five functions of this type of media: decorative, representational, organisational, interpretational and transformational¹⁶. Depending on the type of illustration used, the meaning that observers infer from them could change. In an assessment context, this may affect the behaviours and actions of test-takers. If this can happen within one particular category of media, other types such as animations or simulations should be researched to gain a sound understanding of the psychometric properties of items involving different types of multimedia stimuli. Furthermore, the authors also conceded that building a mental model based on a text may be an important aspect of the construct measured in testing. Thus, 'taking away the need to build mental visualisations might remove that facet from the test and could thereby undermine a test's construct validity' (Lindner et al., 2017a, p. 491).

¹⁶ Decorational pictures 'decorate' the page and bear little meaning to their related text. Representational pictures mirror part or all of the text content. Organisational pictures provide a framework for the text content e.g. family trees. Interpretational pictures help to clarify difficult text by relating complex phenomena to image based analogies e.g. the nucleus of a cell being represented by a police officer directing traffic. Transformational pictures provide mnemonics to facilitate an individual's recall of text information e.g. including black dots in the negative space of the two 'e's in the word 'eye' to recall its spelling (Carney & Levin, 2002, p. 7).

This research suggests that the communication of meaning by test items is closely associated with the use of multimedia objects. The inclusion of images can affect what cognitive processes test-takers engage in. The use of different, more dynamic forms of multimedia, like animations or simulations, may also result in such differences but this is an under-researched area of the literature. Schnotz and Bannert (2003) argued that making sense of, or “processing” visual resources requires the leveraging of prior knowledge to integrate the external representation (text, images etc.) with its internal semantic representation, and thus requires the use of schemas for comprehension. Understanding visual representations involves complex interactions between perceptual surface structures (e.g. key features of the visual resource), deep semantic structures (e.g. an individual’s understanding of these features), and association and inference with cognitive schema (e.g. integrating all relevant knowledge) (Schnotz & Baadte, 2015). When applied to an assessment context, the difficulty of the item with which a multimedia object is associated with would also play a considerable role (Sagoo et al., 2020). Research should aim to understand how different forms of multimedia objects can affect each of these processes to better determine the impact, if any, on construct measurement.

2.6.2.3 Expertise Reversal Effect

The necessity for guidelines relating to the inclusion and design of multimedia items in testing scenarios can also be seen in work by Malone and Brünken (2013). An experimental design was employed by the authors to compare the performance of novice and expert drivers ($n=100$) in a TBA involving multimedia objects (images vs animations) instead of traditional text-based vignettes. A full driving license for more than two years was the criterion used to include individuals in the expert group. Novices were classified as any individual who did not have a full license and was participating in lessons in a designated driving school. Malone and Brünken (2013) found an interaction effect between multimedia type and expertise level. The animated presentation of materials assisted the performance of novices on this assessment. The animations did not allow for the clear identification of expert drivers. The authors demonstrated in their experiment that ‘helpful features’ (Kirschner et al., 2016, p. 24) that increased material coherence in accordance with the CTML interfered with the criterion validity of the test. Kirschner et al. (2016) point to the *expertise reversal effect* (Kalygwa & Renkl, 2010) to explain this phenomenon. The expertise reversal effect constitutes ‘a reversal in the relative

effectiveness of instructional methods as levels of learner knowledge in a domain change' (Kalygna & Renkl, 2010, p. 209). Instructional techniques such as those outlined in the CTML that can maximise the learning of new material can have negative consequences when used with individuals who have already acquired the desired knowledge and skills. While the expertise reversal effect has, at present, only been studied in relation to learning materials, its very existence, alongside the findings of preliminary research by Malone and Brünken (2013), could be used as 'an indicator for the inappropriateness of many design principles for assessment' (Kirschner et al., 2016, p. 21).

Work by Wu et al. (2010) provides another interesting insight in relation to the impact of different multimedia objects on test-taker performance when different levels of prior knowledge are present in TBAs based on Earth Science. The study involved 314 16-year old students in Taiwan where half of the group had completed their studies of the curricular material in the previous semester ($n=194$) and half were still in the process of completing the unit of study ($n=120$). Each group was stratified into three categories depending on their level of prior knowledge (based on three prior school-based summative assessments) – low, medium and high. Using a comparative experimental design where participants completed TBAs that used animated or static stimuli, the authors found that only one of their groups, the group who had completed the unit of work the previous semester, achieved higher results in the animated TBAs ($p=.05$, $d=0.3$). However, a large, practical (rather than a statistically significant) difference was noted in the average scores between animated and static graphic groups when prior knowledge was taken into consideration. It was found that low prior knowledge students performed better in the animated condition ($d=0.7$) while high prior knowledge students performed better when static pictures were used ($d=0.7$). Similar findings were noted by Tai et al. (2006).

Due to concerns over cognitive load, construct measurement and the expertise reversal effect, it appears that the CTML (Mayer, 2008) should not be wholly applied to the design of test items involving multimedia objects. Most worryingly, the improper use of multimedia objects in assessments could lead to Type I and Type II errors. A Type I error (false positive) would mean that a test-taker is incorrectly assigned a high grade or skill level as the use of multimedia objects made the test items easy to answer without the requisite skill levels. Certainly, Lindner et al.'s (2017a) discovery of a multimedia effect in testing suggests that this is certainly possible. Alternatively, Type II errors (false

negative) are also possible as a test-taker who does have the required knowledge and skills is not considered to have them. Work by Malone and Brünken (2013) supports this possibility. Based on the evidence presented here, it is not necessarily appropriate to apply the principles and assumptions of the CTML (Mayer, 2008) to assessments without some element of modification. This has led to the emergence of the Cognitive Theory of Multimedia Assessment (CTMA; Kirschner et al., 2016).

2.6.3 A Cognitive Theory of Multimedia Assessment (CTMA)

Kirschner et al. (2016) claimed that the ‘design principles derived from the CTML need to be varied or even reversed’ for assessment scenarios (p. 20). Instead of using these design principles to minimise the cognitive load associated with learning, they should instead be modified to achieve the *optimal* level of *cognitive assessment load* necessary to discriminate between test-takers with varying levels of ability and skills. This aim is the foundation of Kirschner et al.’s (2016) proposed CTMA. Kirschner et al. (2016), proposed a triarchic model of cognitive *assessment* load, influenced by the work of DeLeeuw and Mayer (2008). Table 2.3 summarises the different forms of cognitive assessment load. The authors argued that varying these forms of assessment loads can ensure that a balance between ecological validity and criterion validity can be achieved.

Table 2.3 Forms of cognitive assessment load

Assessment Load Type	Description
Intrinsic	This is the assessment load that arises from the subjective difficulty and complexity of a task.
Extraneous	An easily manipulated form of assessment load associated with the inclusion of incoherent or unnecessary information.
Germane	A form of assessment load that is produced by the process of information retrieval and problem solving.

As discussed in Section 2.6.2.3, the expertise reversal effect (Kalygua & Renkl, 2010), whereby techniques that support novices’ learning can interfere with an expert’s performance, suggests that an uncritical adoption of multimedia design principles to non-

learning scenarios could threaten the criterion validity of an assessment (Kirschner et al., 2016). According to the APA (2019), criterion validity is 'an index of how well a test correlates with an established standard of comparison'. Therefore, if someone is deemed to be an expert, they should perform much better on a test than those who are not experts. Kirschner et al. (2016) argued that in an assessment situation, this expertise related difference in performance is caused by 'different amounts of intrinsic and germane cognitive load in experts as compared to novices' (p. 22). Experts, due to having more prior knowledge in the area, have a lower level of intrinsic assessment load in a task than novices. Furthermore, an expert's ability to manage germane assessment load is higher than a novice's as their information retrieval and problem solving skills are more developed. When measuring most constructs, items in assessments should reveal these different assessment load capabilities in test-takers. A novice's intrinsic assessment load in relation to a particular item should not be decreased by the form of stimuli used nor should the design of an item mask an expert's higher levels of germane assessment load.

To ensure that an item can accurately determine a test-taker's intrinsic and germane assessment load (which should vary according to the task and the individual's expertise level), an optimal level of extraneous assessment load should be introduced to the task or item (Kirschner et al., 2016). As experts should be able to perform in suboptimal circumstances (Kalygna & Renkl, 2010), managing the amount of extraneous assessment load (which is caused by unnecessary or redundant information) within an item should allow for the differentiation of different levels of expertise. With an optimal amount of extraneous load 'experts should still have free resources to accomplish the tasks while the novices' complete cognitive capacity will be consumed by intrinsic and extraneous load' thus making them less likely to perform well (Kirschner et al., 2016, p. 23). To determine the limits of a test-taker's knowledge, skills and abilities, tasks with increasing levels of cognitive assessment load should therefore be present in an assessment as demonstrated in Figure 2.10. An optimal level of extraneous load is best achieved by ensuring that the item corresponds with the real-world situation. Items can be made more or less difficult by adding or removing features that control this level of extraneous assessment load.

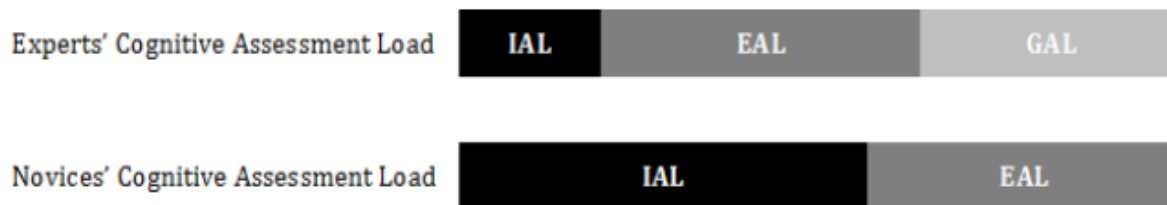


Figure 2.10 Intrinsic, Extraneous and Germane Assessment Load for experts and novices for a difficult task (Reprinted from Kirschner et al., 2016, p. 23)

The key preposition of the CTMA is that cognitive assessment load should be carefully managed but not eliminated. An optimal level of extraneous assessment load can be achieved by modifying the principles of the CTML (Mayer, 2009). Table 2.4 outlines Kirschner et al.'s (2016) recommended adaptations for the CTML.

Table 2.4 Adapting the CTML for Assessment (adapted from Kirschner et al., 2016, p. 30)

CTML Design Principle	Adaptation for Assessment
<i>Coherence:</i> Irrelevant information should be excluded	The amount of coherence in the item should reflect the coherence found in the 'real world'. Modifying this will vary the level of extraneous assessment load.
<i>Signalling:</i> Signals should guide the learner	The removal of additional cues can increase extraneous assessment load.
<i>Redundancy:</i> Audio narrations should not be accompanied by on screen text	The amount of redundant information from any channel should represent the levels that would be seen in 'the real world'. Adjusting the amount of redundancy in an item can modify the level of extraneous assessment load.
<i>Spatial Contiguity:</i> Images and labels should be presented side by side	The spatial contiguity of the materials within the item should reflect the real world to ensure optimal extraneous assessment load.
<i>Temporal Contiguity:</i> Audio narrations should be synced with each segment or event	The temporal contiguity of the materials within the item should reflect the real world to ensure optimal extraneous assessment load.
<i>Segmenting:</i> Information should be presented in short 'chunks' or segments.	The task should be presented as a continuous unit to ensure that an individual's level of germane assessment load is apparent.
<i>Pre-training:</i> Key words should be presented to learners before learning occurs	No pre-training should be given in assessment contexts to ensure that an individual's intrinsic assessment load is consistent with their ability.
<i>Modality:</i> Text based information should be conveyed by aurally.	The presentation of information using modes that are most reflective of the real world will produce an optimal level of extraneous load.
<i>Multimedia:</i> Words and pictures should be together.	Pictures should only be used where appropriate.
<i>Personalisation:</i> Informal language should be used.	Experts should be able to compensate when formal or informal language is used whereas novices tend to prefer informal language.

2.6.3.1 Evaluating the CTMA

The CTMA presents a strong argument in favour of an independent cognitive theory of multimedia assessment that does not blindly adopt assumptions from learning theory. Instead, Kirschner et al. (2016) have made a concentrated effort to build a new assessment-focused theory based on sound principles from cognitive psychology. In doing so, it provides a research agenda for the field, encouraging the design of studies that can support or reject their proposed design principles. Kirschner et al.'s (2016) work represents a much needed reconceptualisation of the purpose and value of multimedia materials as they apply to the field of assessment. However, the theory is still in the initial stages of development and requires further work and refinement.

For example, Kirschner et al.'s (2016) theory discussed the role of multimedia objects in assessment materials. The arguments that they use to support their ideas come from the use of multimedia objects as item stimuli (e.g. Malone & Brünken, 2013). Yet, technology-based items involve two components: the *stimulus piece* and the *interaction space*. The item's interaction space is where the test-taker's actions and responses are recorded and this can encompass a wide variety of options including drag-and-drop, line and object production and the upload of sound, image and video files (Russell, 2016). The CTMA acknowledges that the correct design of both item components is important and claims that interaction spaces should contain response actions that are representative of the particular domain or skill being assessed. The authors asserted that there should be a high degree of fidelity between what the test-taker is asked to *do* in the assessment and what would occur in a real-life application of the construct being measured. This point is particularly relevant for items and tasks involving interactive simulations.

Despite Kirschner et al.'s (2016) assertions that response actions need to be considered when designing tasks in TBAs that involve multimedia objects, very little time is spent discussing this aspect of assessment in either practical or theoretical terms. The CTMA, in its current form, does not make sufficient effort to integrate the role of response actions into its theory, despite it being a critical element of technology-based items. While it is likely that the type of response action associated with an item would have an impact on the germane assessment load of test-takers, this connection was not made explicit or even developed within the theory. Given the crucial role this part of an item plays in the assessment of a test-taker's knowledge, skills and abilities, this is a significant omission. Furthermore, the design principles proposed by the CTMA do not offer any best practice

guidelines that could support the selection of appropriate response actions that would ensure optimal assessment load. Therefore, it can be argued that the CTMA cannot be considered a complete theory of multimedia assessment as it does not give adequate attention to one-half of an item's features.

Expertise research has heavily influenced the CTMA. As a result, this theory is particularly concerned with criterion validity, where a person's level of expertise or skill according to another measure should still be evident in the assessment involving multimedia materials. The design principles suggested by the CTMA aim to optimise the discrimination between novices and experts in assessments. While it is important to ensure that assessments can effectively distinguish between experts and novices, the impact of multimedia objects on other forms of validity should also be considered when formulating such design principles. As demonstrated by other research studies (e.g. Malone & Brünken, 2013) the type of multimedia object used in an assessment could vary the construct being measured or even the behaviours that test-takers engage in (e.g. Vorstenbosch et al., 2014). Therefore, the design principles of the CTMA could be considered incomplete as they do not fully take into consideration the possible relationship between multimedia objects and construct validity. To advance the field of TBAs, research in this area is required.

2.7 Advancing the field of TBAs

The transition to TBAs has prompted the development of a range of new item types with more complex perceptual elements (e.g. multimedia objects) or interaction spaces (e.g. drag-and-drop). These alternative item formats may potentially provide additional information about test-taker reasoning or cognitive skills, thus strengthening the measurement of a particular construct. As seen in the previous discussion on the application of the CTML to TBAs, the design of these items is not being informed by appropriate research that fully aligns with testing principles. Failure to have a fully informed understanding of how technology-based items should be designed, taking into consideration their use of multimedia objects and more complex interaction spaces, could negatively impact the validity of TBAs. In an attempt to address this issue, Moon et al. (2019) noted that answering the 'question of how item formats affect test-taker cognition' should be prioritised in research (p. 54). This will facilitate the interpretation of test scores and help test-developers understand if score differences are due to

inappropriate item design or test-taker response characteristics associated with item formats e.g. drag-and-drop items taking longer to complete, items with animations being easier etc. This may then allow test and item developers to make better design decisions. Recent work by Moon et al. (2019) would certainly support the value of research that explores item design features at a ‘fine-grained level rather than making a sweeping assumption for a give type’ (p. 61). Their between-groups experimental study involving 1091 adults completing content equivalent questions in mathematics with many different interaction spaces found that test-taker response tendencies across different item formats could potentially affect test scores and their psychometric properties.

Understanding the possible impact that a multimedia object or a particular interaction may have on the assessment of a particular construct will be essential if TBAs are to positively contribute to educational assessment. Achieving this understanding is best done through the lens of the Evidence-Centred Design (ECD) conceptual assessment framework (Mislevy et al., 2003). This is a construct-centred approach to designing assessments which asks ‘what complex knowledge, skills or attributes should be assessed... what behaviours should reveal those constructs and what tasks or situations should elicit those behaviours’ (Messick, 1994, p. 16). ECD is a coherent integration of three main components (summarised in Figure 2.11):

- *Proficiency Models* – The proficiency model focuses on describing the test-taker in terms of the constructs that are to be measured. These descriptions often include guidelines in terms of the proportion of a domain of tasks that students are likely to answer correctly. It aims to set out *what* is being measured.
- *Evidence Models* – This model focuses on the very nature and the recommended analysis of the responses, identifying *how* the construct should be measured. It aims to define observable variables and indicators of performance that should be interpreted to give guidance on a test-taker’s level of proficiency. According to Hao and Mislevy (2018), the evidence model is composed of two parts. The measurement model identifies the rules and procedures that characterises and grades different work products as specific numeric or symbolic values. These are the observable variables. The evaluation component refers to the psychometric models that are used to combine data from the observable variables to provide information on a test-taker’s proficiency.

- *Task Models* – The task model describes ‘how to structure the kinds of situations needed [sic] in order to obtain the kinds of evidence necessary for the evidence models’ (Groff, 2018, p. 193). It delineates *where* the construct can be measured and it is here that the items of an assessment are described. It also designates how the materials that should be presented to the test-takers and what work products to be generated using particular response options. Recent research on task models have expanded this model to include two key components – work products and stimuli (Hao & Mislevy, 2018).

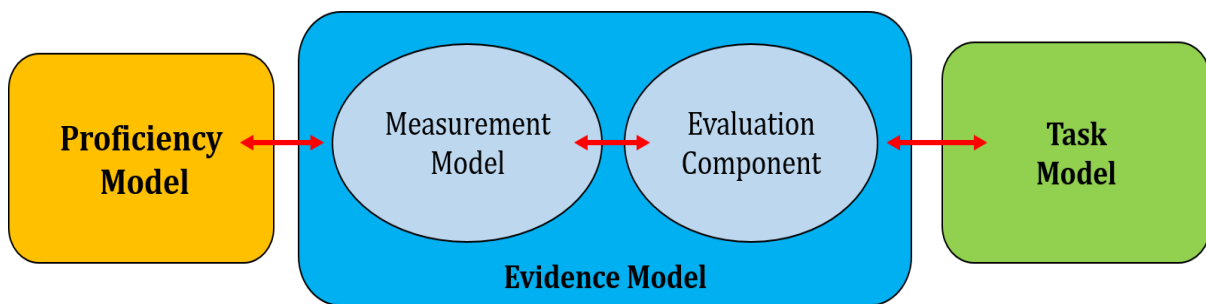


Figure 2.11 Evidence Centred Design (ECD) conceptual assessment framework (modified from Hao & Mislevy, 2018)

As noted by Groff (2018), the ECD framework offers some much needed rigour and coherence to the design of technology-based items and TBAs. Unfortunately, it appears that there is limited guidance about how to link the assessment of complex constructs with the design of assessment tasks in accordance with the procedures associated with the Task Model part of this framework (Arieli-Attali et al., 2019). As outlined by Mislevy et al. (1999), the *Task Model* provides a framework for describing the situation in which test-takers can demonstrate their knowledge, skills and abilities. This framework includes the ‘stimulus materials, conditions and affordances... [and] specifications for the work product’ (Mislevy et al., 1999, p. 19). As noted by Arieli-Attali et al. (2019), variables that influence task difficulty, task management and presentation all need to be taken into account in the design of the Task Model. It is here that the work of Kirschner et al. (2016) and the CTMA can add true value to the field of TBAs. The CTMA aids the process of developing guidelines regarding the use of stimuli to help design appropriate tasks involving multimedia objects in TBAs, particularly in relation to task difficulty and presentation. The CTMA can offer some much needed support to design of

assessment tasks in accordance with the ECD framework. Similarly, the ECD framework can allow the CTMA to be applied in a way that fully takes into consideration the constructs that are being assessed.

What is included in the Task Model of an item is influenced by the definitions of the construct outlined in the Proficiency Model and the psychometric procedures required by the Evidence Model. In this way, the characteristics of an item are ‘determined by the nature of the behaviours that provide evidence’ for the particular construct (Arieli-Attali et al., 2019, p. 9). The ‘nature of the behaviours’ associated with the targeted construct are reflected in the work products produced by the test-taker. The work products are created in the ‘interaction space’ of technology based items and, as previously discussed, can include a range of response actions. Research should explore what type of evidence is collected from different response actions or items involving different multimedia stimuli. It should also consider whether this evidence fully supports what is being described in the Proficiency Model. This research would significantly enhance the CTMA and the ECD framework.

The basic assumptions and design principles of the CTML have been adjusted to reflect the aims and purpose of assessment in the CTMA. For that reason, it is unsurprising that the CTMA seems, according to its own authors, unfinished. By focussing on the differences between assessment and learning only, it presents a limited view of what should be considered when developing high-quality tests and assessments. By associating it with the ECD framework, the CTMA can be a part of a more coherent approach to the design of technology-based items that has a construct-driven approach to the assessment process, rather than one that is characterised by a ‘techno-centric’ mindset or that focuses too heavily on criterion validity. A variety of approaches should be used to obtain the information necessary to allow the CTMA to support the development of appropriate task models that will enhance the quality of assessments designed within the ECD framework. These include the use of eye-tracking equipment to collect response process data.

2.7.1 Value of Response Process Data

Embretson (2016) rightly noted that the basis of examinees’ responses to items (i.e. their thinking processes and their actions) is an important aspect of validity. It also underpins the ECD Framework’s Evidence Model. If a test is to be considered valid,

examinees' responses to test items are expected to reflect the construct being measured. To investigate this, response process data collected from examinees can be inspected as it represents 'the mechanisms that underlie what people do, think, or feel when interacting with, and responding to, the item or task and are responsible for generating observed score variation' (Hubley & Zumbo, 2017, p. 2). Examination of these data can often reveal very interesting insights about a test item that may influence how test scores are interpreted. For example, if a test-taker correctly selects an answer in a reading comprehension test, test specifications often assume that the test-taker read the required passages and selected the relevant information needed to answer the test item. However, when research on test-taker response processes is available validity issues usually emerge. For example, in their eye-tracking study examining a reading test involving 'fill in the gap' items, Paulson and Henry (2002) found that these items caused readers to radically alter their reading process in order to complete the assessment successfully. The test-takers did not engage in many of the reading behaviours outlined in the test's specifications. The authors concluded that this test could not be considered an 'accurate measure, or even a modest approximation of, the reading comprehension process' that it described (Paulson & Henry, 2001, p. 242). Therefore, to preserve the validity of the interpretations being made from a test, and to ensure that test-takers' behaviours in an item's interaction space are consistent with what the item is assumed to measure, response process data should be collected. Eye movement data as a form of response process data may be particularly useful for the field of digital assessments.

2.7.1.1 Eye-tracking Technology and Eye Movement Data

Eye-tracking technology can be used to gather response process data and explore how people process information in TBAs involving different multimedia tasks and activities. Eye-tracking refers to 'a set of technologies that make it possible to establish the eye gaze of an individual' (Navarro et al., 2015, p. 2237). By using infrared beams that are reflected onto an individual's pupils and then recorded, Hyöna (2010) noted that eye-tracking technology can allow researchers to identify what is attended to first in a presentation, and for how long, along with other information related to the attentional allocation processes of humans. Researchers can programme eye-tracking technology to calculate a range of measures for specific Areas of Interest (AOIs) in presentations. Two main types of measurements are obtained using eye-trackers: fixations and saccades. Just

and Carpenter's (1980) eye-mind hypothesis asserted that eye fixations, which 'describe the stable state of the eye at one point' (Alemdag & Cagiltay, 2018, p. 414), reflect the attention process. Similarly, saccade measurements represent the eye movement between fixations, which shows the change in the focus of visual attention. Lai et al. (2013) categorised eye-tracking data into three scales – temporal, spatial and count. Temporal scales relate to time spent in specific eye movements e.g. total fixation duration, time to first fixation etc. Lai et al. (2013) defined spatial scales as 'locations, distances, directions, sequences, transactions, spatial arrangement or relationships of fixations or saccades' e.g. saccade lengths, fixation sequence (p. 93). Count scales represent the frequency of specific eye movements e.g. total fixation count. Each of these scales are thought to help researchers make more direct inferences about the cognitive activities of an individual e.g. what information is considered important, what information is frequently referred to etc.

Alemdag and Cagiltay's (2018) systematic review of 58 eye-tracking studies (the majority of which involved college students with science learning materials) that explored multimedia learning found that there was sufficient research to support the association between different eye movement measures and a range of cognitive processes. For example, a large amount of research supported the association between different eye-tracking metrics and the cognitive processes of selection (e.g. time to first fixation on AOIs), organisation (e.g. total fixation count on an AOI) and integration (e.g. saccade scan path analysis). Work by Zu et al. (2018) showed that mean fixation duration and AOI transitions were sensitive to measuring different types of cognitive load. Alemdag and Cagiltay (2018) also noted that there was evidence to support the relationship between certain eye movement measures and learner performance. This included the positive association between visual search efficiency with learning. In concluding their review of the literature, Alemdag and Cagiltay (2018) called for more research exploring multimedia information processing of non-college age students with non-science based learning materials. Even addressing one of these issues would be a notable contribution to the field. The authors also recommended that more eye-tracking research should use the spatial scales of eye-tracking measures. Fixation position, fixation sequence, and scan path patterns can show spatial sequences of visual attention over time in detail. Only a limited number of studies which examine these metrics exist. Interestingly, Alemdag and Cagiltay (2018) failed to identify one of the most significant

omissions surrounding the use of eye-tracking technology to explore multimedia materials. Eye-tracking technology to explore the use of multimedia materials in *assessment contexts* is less common despite the fact that TBAs are becoming more common.

Multimedia materials are frequently integrated into test items in large-scale assessments (e.g. PISA: OECD, 2018; 2016a; 2016b), yet little is known about how multimedia elements affect cognitive processing in item solving (Lindner et al., 2017a). Eye-tracking research within the field of multimedia learning has revealed that certain design decisions related to the use of multimedia objects can support or hinder learning. For example, students in the Wang et al. (2016) study, exhibited significant difficulties learning a new recipe when they had to co-ordinate incoming information that used a variety of formats (e.g. text and video). Wang et al.'s (2016) eye-tracking study found that high inter-scanning counts (the number of times the eye moved between two AOs) between text and video information was a negative predictor of recall performance. Given the increasing use of multimedia materials in TBAs, it is important to understand that if a multimedia object can affect learner performance, the same may be true of test-taker performance. In testing situations, test-takers must understand the problem that is presented in the test item stem to be able to solve the item correctly. Lindner et al. (2017a) argued that the deployment of multimedia materials in assessment materials may influence the development of such mental models by test-takers.

As demonstrated from the research conducted using learning materials, eye-tracking is a suitable method for revealing the cognitive processes undertaken by test-takers in TBAs to create those mental models. Eye-tracking research may be particularly well suited to TBA research as the components of many technology-based items are often displayed in distinct locations, thus allowing for clear spatial metrics to be obtained to reveal how test-takers solve an item in a TBA. Using metrics such as these, Lindner et al. (2017b) found that the item-solving process in TBAs can be roughly divided into two phases: (1) an information-acquisition phase, in which students construct a mental representation of the problem or situation and (2) a decision-making phase where the answer options are evaluated before a final choice is decided upon. As Lindner et al.'s (2017a; 2017b) work was one of the first to address this issue research such as this should be repeated to determine the replicability of the results. Furthermore, Lindner et al. (2017a; 2017b) found evidence in favour of a multimedia effect in testing, similar to

the multimedia effect in learning, which is the basis of the CTML. Lindner et al. (2017b) found a significant decrease in item difficulty when items using text and pictures as stimuli were compared to corresponding text-only items ($d=0.66$). Test-items with pictures facilitated the construction of a mental model of the problem task as indicated by the reduced reading time of the item stem. Test items involving pictures also allowed test-takers to be quicker at making decisions and dismissing distractors in MC type questions ($r=-.59$). Lindner et al. (2017b) also found that test-takers directed their attention to the picture rather than the text in the initial information-acquisition phase and also in the early decision-making phase.

Eye-tracking research can provide some crucial information about how test-takers engage with TBAs when multimedia materials are included. Research exploring how test-takers engage in TBAs involving multimedia objects would provide important information about the impact of certain design decisions on test-taker attentional allocation behaviour and performance in TBAs. Further research is required to fully understand how multimedia objects interact with test-taker's cognitive processes in assessment contexts to ensure that valid and appropriate interpretations are being applied from the research to the field of TBAs. For example, Lindner et al. (2017b, p. 101), claimed that the multimedia effect was a 'welcome change' as it allowed test-takers to become more efficient information processors. However, this facilitative effect may not be desirable in test-taking situations where test-takers are required to process information under sub-optimal conditions in order to demonstrate their proficiency. Future eye-tracking research within the field of TBAs needs to be informed by appropriate theoretical frameworks such as the CTMA and the ECD framework in order to best understand the research findings. These findings should emerge from experimental research.

2.8 Summary and Conclusions

As multimedia objects and stimuli have been widely used for teaching and learning, educational research has been able to identify various design parameters of effective multimedia objects for learning (e.g. Mayer, 2009). Yet, the optimal design and deployment of multimedia stimuli for test items in assessment contexts has not been researched in a similarly systematic and coherent manner. In particular, one of the most important aspects of a multimedia object that should be considered when including them

in TBAs is its modality, which can be static (e.g. images), dynamic (e.g. animations, videos) or interactive (simulations). As demonstrated by Malone and Brünken (2013), static and dynamic representations of situations can interact with a range of test-taker characteristics, including expertise, to influence test-taker performance. Work by Quellmalz et al. (2013) demonstrated that simulations can change the construct being measured and the behaviours of test-takers. Consequently, instead of assuming that one modality is always better than another, researchers should address the complex factors that could influence the effectiveness of different multimedia representations for different populations and testing purposes. Studying the impact of different multimedia objects on test-taker performance and attentional behaviour will allow them to be used in 'a more targeted manner and based on empirical findings rather than on the individual theories of test constructors' (Lindner et al., 2018, p. 376).

Exploring the impact of different modalities on test-taker performance is best achieved through experimental research where modalities can be controlled and manipulated to infer causal conclusions about their efficacy which future researchers can replicate. Unfortunately, with the exceptions of Quellmalz et al. (2013) and Wu et al. (2010), very few experimental studies have addressed how different types of representations can affect test-takers' performance. This should be addressed as soon as possible to help inform the design of future technology-based items. Test-taker behaviour should also be considered in any future research that aims to inform item writing guidelines. Understanding *how* test-takers engage with test items that use different forms of multimedia objects can provide key insights into the cognitive processes that underlie test-takers' actions in a test. Collection of process data would allow test-developers to determine if certain multimedia objects allow test-takers to engage, or not engage, in behaviours that align better with the construct they aim to measure. A greater body of experimental research that involves process data in some way needs to be available in the field of TBAs so that multimedia stimuli that can support the measurement of a targeted construct can be deployed appropriately.

In summary, this literature review has revealed significant gaps in knowledge, thus justifying the study outlined in Chapter 3. The study was designed with the aim of examining the impact of static, dynamic and interactive multimedia stimuli on test-taker performance and attentional behaviour across a range of item types. The proposed study will involve an exploration of the influence of different item designs (in terms of item

stimuli and interaction spaces) in TBAs on test-taker performance using an experimental approach. Response process data in the form of eye movements will also be collected to allow inferences on underlying cognitive processes to be drawn, which may help to explain if and why certain multimedia stimuli or interaction spaces can influence item characteristics or test-taker performance. The specific research questions that guided the design of this study are outlined in the following chapter.

Chapter 3

Methodology

3.1 Introduction

This study explored the design of technology-based items that incorporated multimedia stimuli (e.g. images, animations, simulations) and interaction spaces that contained response actions which varied in constraint (SR, CR, FR). Using an experimental approach appropriate for educational research (Coleman, 2019), the study examined how different multimedia objects can influence a test item's capacity to accurately measure the targeted knowledge, skill or ability of a test-taker. For the purposes of this study, two versions of the same digital assessment were compared. One version employed a standard text-image paradigm for the presentation of item stimuli. The other used animations. The differences in test-taker performance and behaviour were examined using multiple data sources, specifically test score data, eye movement data and interview data. Test-taker performance and attentional behaviour in relation to simulation-type items were also investigated in this study. This chapter describes the methodology that was used to address the aims of the study. To begin, the conceptual framework and research questions for the study are outlined. The research design and sampling procedures are then described, along with an overview of the instruments that were used in this study. A description of the key measures and variables of the study, as well the associated ethical considerations, are also presented.

3.2 Conceptual Framework

According to Maxwell (2005), the purpose of a conceptual framework is to present 'the system of concepts, assumptions, expectations, beliefs and theories that supports and informs the research' (p. 33). Eisenhart (1991) asserted that a conceptual framework should also justify the issues chosen for investigation. The framework depicted in Figure 3.1 is intended to summarise how the issues highlighted in the literature review informed the design of the current study

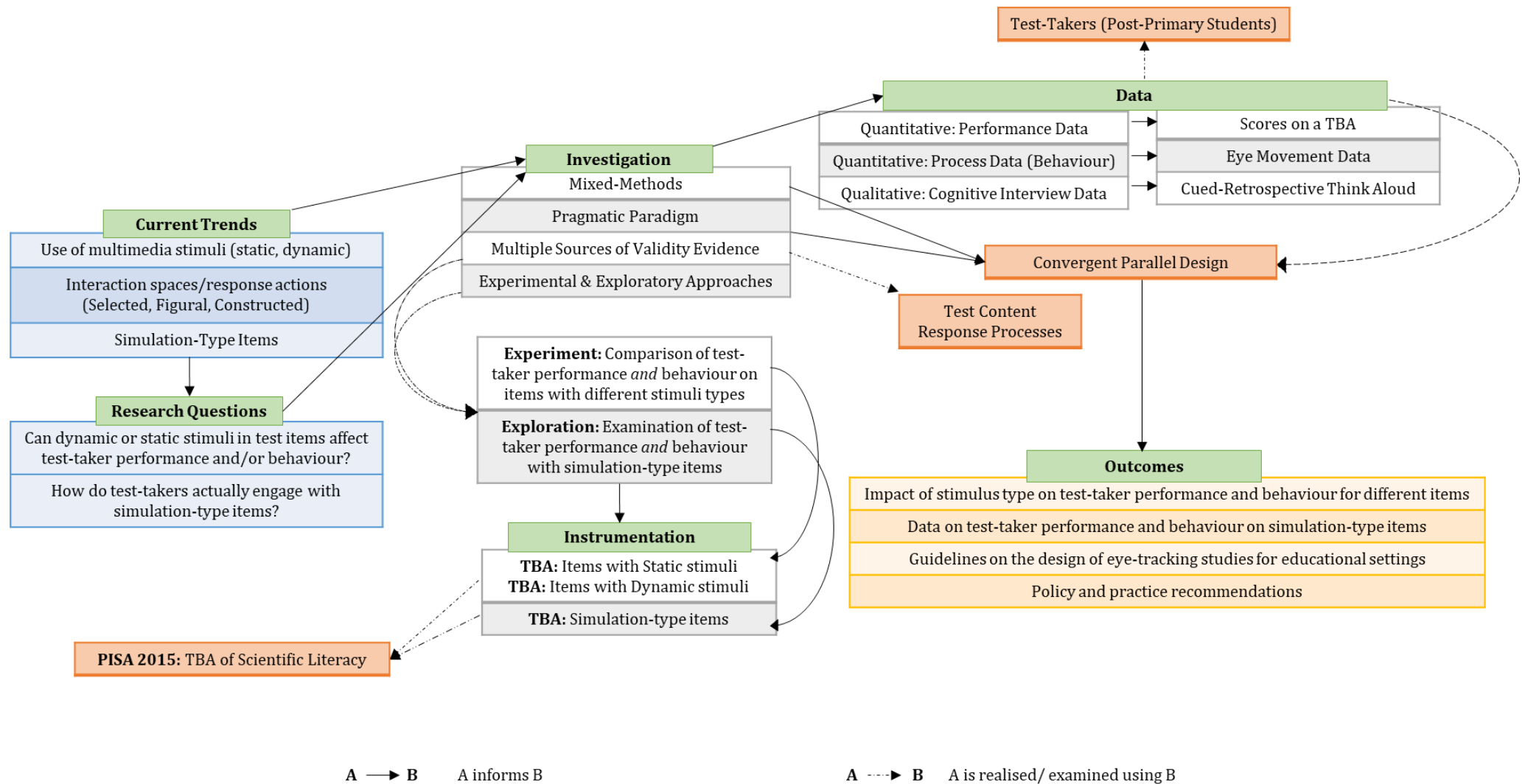


Figure 3.1 Conceptual framework underlying the current study

Three '*Current Trends*' in the field of digital assessments and TBAs were identified in the conceptual framework based on the literature reviewed – the use of multimedia stimuli, the range of possible response actions and the emergence of interactive multimedia test items involving simulations. However, as discussed in Chapter 2, these advances in the field of TBAs have not been fully researched. For each of these developments, the conceptual framework extracts one unanswered question from the literature under '*Research Questions*'. For example, the value of using animations in assessment contexts when compared with static images has been under-researched. As a result, it is unclear when static or dynamic multimedia objects should be used in test items. While it would have been beyond the scope of this thesis to provide definitive answers to any of the questions presented in the conceptual framework, the study that arose from this framework aimed to contribute to the relevant knowledge needed to begin to answer such questions.

The conceptual framework represented in Figure 3.1 also outlines how such questions were addressed in the current research under the heading '*Investigation*'. A mixed methods approach anchored by a pragmatic paradigm was applied. This paradigm advocates a plurality of research approaches to address and answer the complexity of real-world research (Creswell, 2014; Onwuegbuzie & Leech, 2005). Collection of validity evidence was also considered necessary to address the stated research questions. While the *Standards* (AERA et al., 2014) include five sources of validity evidence in total, researchers such as Embretson (2016) noted that validity evidence 'typically does not include all five aspects and evidence for the response processes aspect is often not included' (p. 7). In fact, Embretson (2016) acknowledged that for educational achievement tests evidence for the test content aspect dominates the field. Yet multiple sources of validity evidence are necessary if there is to be any confidence in the design of items in TBAs. Consequently, the collection of validity evidence involving 'Test Content' and 'Response Processes' was a priority in the current investigation.

This validity evidence was acquired using experimental and exploratory approaches. Experimental studies are considered the most effective way of identifying possible causal relationships between variables (Fraenkel & Wallen, 2006). They are the preferred methodology in determining differences between test-takers' actions according to item design e.g. use of multimedia stimuli. Yet, the value of exploratory studies for the field of TBAs should also be emphasised. For example, research to date has

not communicated a clear understanding of how test-takers interact with more complex items (e.g. simulations) and if such interactions can better represent the constructs being measured by an item. Exploratory studies can provide further information on both these issues and may then contribute to the development of more effective experimental research in the future. Therefore, both approaches were considered necessary based on the research interests of the study. Different TBAs however, are needed for each approach (see '*Instrumentation*', Figure 3.1). The TBA of scientific literacy designed by the OECD (2017) for PISA 2015 provided the instruments required for both the experimental and exploratory studies involved in this research.

The key features of the proposed study, as outlined by the '*Investigation*' box in Figure 3.1, informed what '*Data*' were collected. As a mixed methods approach was considered the most appropriate, quantitative and qualitative data were required. Given the study's interest in test-taker performance *and* behaviour in relation to different test items and the necessity for multiple sources of validity evidence, two sources of quantitative data were obtained: test-taker scores on a TBA ('Test Content') and a numerical summary of participants' eye movements ('Response Process') while completing said TBA. Qualitative data were gained using a form of cognitive interviewing associated with eye movement research – a cued-Retrospective Think Aloud (Elbabour et al., 2017). As discussed previously, secondary school students are an under-researched group in relation to the design and use of TBAs. Their inclusion in the study addressed a significant shortcoming in the literature. Using a convergent parallel design, these data were analysed and interpreted to provide a number of study '*Outcomes*'. By identifying the impact of stimulus type on test-taker performance and behaviour for different items and by gathering further data on test-taker engagement with simulation-type items, the research approach outlined in this conceptual framework was able to provide essential recommendations on the design and use of items in TBAs for second-level students. Guidelines on the design of eye-tracking studies for educational settings also emerged as a result.

The conceptual framework outlined in Figure 3.1 identifies the components of an effective study that can further knowledge on the design and use of items in TBAs. This chapter will now outline how these considerations informed the current study, beginning first with the research questions.

3.3 Research Questions

Based on the gaps in knowledge identified from reviewing the relevant research literature, four main research questions were developed to guide the study. Each of these questions is presented below, in addition to a number of linked sub-questions.

Research Question (RQ) 1: Do different multimedia stimuli (e.g. images, animations) affect test-taker *performance* in TBAs?

RQ1a: Is the performance of test-takers on items in a TBA affected by the type of multimedia stimulus used?

RQ1b: Is the performance of test-takers on items in a TBA affected by the type of multimedia stimulus used when their previous levels of knowledge are considered?

RQ1c: Does the type of multimedia stimulus used affect key item statistics (i.e. difficulty, discrimination)?

RQ2: How do different multimedia stimuli (e.g. images, animations) affect the *attentional behaviour* of test-takers in TBAs?

RQ2a: Does the *number of visits* to an item's interaction space differ according to the multimedia object used?

RQ2b: Does the *average duration of whole fixations* in the interaction space of an item differ according to the multimedia stimulus used?

RQ2c: Does the *proportion of fixations* in relation to the interaction space of an item differ according to the multimedia stimulus used?

RQ3: What behaviours are demonstrated by test-takers when responding to items and tasks involving *simulations* and are any of these related to overall performance?

RQ3a: Are time-on-task and time-per-phase related to test-taker performance?

RQ3b: What relationship, if any, does the number of simulations run per task have on test-taker performance?

RQ3c: What attentional behaviours (*number of visits*) do test-takers exhibit when completing simulation-type items in the Orientation phase?

RQ3d: What attentional behaviours (*time-to-first-fixation, number of whole fixations, proportion of fixations*) do test-takers exhibit when completing simulation type items in the Output phase? Do these behaviours differ by performance?

RQ4: What thought processes underlie test-takers' interactions with items in TBAs?

To address these research questions, three related studies involving a convenience sample of post-primary students (aged between 15 and 17 years) were conducted. The participants came from six schools across Ireland. Study 1A ($n=251$) involved an experimental comparison where participants were randomly assigned to one of two groups. In the control group, participants were presented with items in a TBA that used static images and text as their multimedia stimulus. In the experimental group, participants saw narrated animations for their item stimuli. The data from Study 1A addresses the first research question (**RQ1**). Study 1B involved 33 participants who completed the TBA in Study 1A. Eye movement data were also collected from this group of participants during Study 1B (**RQ2**). Study 2 ($n=24$) involved those participants who completed Study 1B engaging with an additional five simulation-type items (**RQ3**). Performance and eye movement data were collected from these participants. A further subset of this sample ($n=12$) also participated in a cognitive interview to provide additional information to aid in the interpretation of this eye movement data (Study 3). This will provide insight into **RQ4** and will also provide further information necessary to address **RQ1**, **RQ2** and **RQ3**. Figure 3.2 provides a graphical summary of the studies that were implemented to answer the stated research questions.

Participants	Convenience sample of second-level students (aged 15-17 years)
Instrument	TBA with SR, FR and CR items
Study 1 (A, B) <i>n</i> =251, <i>n</i> =33	
Group 1 Experimental (Dynamic: Animations) Group 2 Control (Static: Text + Image) <ul style="list-style-type: none"> • Test scores as a measure of performance (Study 1A, Study 1B) • Eye movement data as a measure of attentional behaviour (Study 1B) 	
Study 2 <i>n</i> =24	
Five simulation-type items <ul style="list-style-type: none"> • Test scores as a measure of performance • Eye movement data as a measure of attentional behaviour 	
Study 3 <i>n</i> =12	
Cognitive interview with participants involved in Study 1B and Study 2	

Figure 3.2 Outline of Study 1 (A, B), Study 2 and Study 3

The remainder of this chapter will now provide more detail on the research design (*Section 3.4*), followed by more specific information on the sampling techniques (*Section 3.5*). The TBA used for this study will be described in *Section 3.6*. *Section 3.7* will summarise the equipment used in the studies. A summary of the eye movement measures and cognitive interview protocols that were used will be outlined in *Section 3.8*. Information about on the pilot study (*Section 3.9*) and main study (*Section 3.10*) will then be provided along with the ethical considerations for this research (*Section 3.11*) and details of the data analysis procedures (*Section 3.12*).

3.4 Research Design

3.4.1 Study 1 and Study 3: Mixed Methods Factorial Design

‘Mixed methods’ is a research approach whereby both quantitative and qualitative data are collected and analysed within the same study. Johnson et al. (2007) characterised it as the combination of ‘qualitative and quantitative research approaches (e.g. use of qualitative and quantitative data collection, analysis, inference techniques) for the broad purposes of breadth and depth of understanding and corroboration’ (p. 123). The purposeful integration of both forms of data allows researchers to obtain a more comprehensive answer to their research questions as they can view phenomena from different viewpoints and through diverse research lenses (Shorten & Smith, 2017). As a result, a mixed methods approach was considered to be particularly relevant to the current study as the collection and analysis of quantitative and qualitative data would better resolve the research questions than either approach alone. Scores on a test of scientific literacy and relevant eye-movement data from participants represented the quantitative data collected in the study. The qualitative data were derived from participant responses to a cognitive interview.

A convergent parallel design was used to collect, analyse and interpret the quantitative and qualitative data. A convergent parallel design involves the researcher concurrently conducting the quantitative and qualitative elements of the study in the same phase of the research process, weighting the methods equally, analysing the two components independently, and then interpreting the results together (Creswell & Plano-Clark, 2011). The use of a convergent parallel design allows the researcher to *explain* rather than just *describe* test-taker interactions with different types of items. This reflects a pragmatic¹⁷ philosophical approach to educational research. This research process is summarised in Figure 3.3.

¹⁷ Pragmatism is a paradigm that advocates a relational epistemology, whereby relationships between constructs are best determined by ‘what the researcher deems appropriate to that particular study’ (Kivunja & Kuyini, 2017, p. 35). It also admits that there is no single reality and that research into these multiple realities should ultimately benefit society (Kivunja & Kuyini, 2017). Mixed methods methodologies are prioritised under this paradigm as the ‘plurality’ of approaches involved allows researchers to choose a combination of quantitative and qualitative methods to answer their research questions (Onwuegbuzie & Leech, 2005).

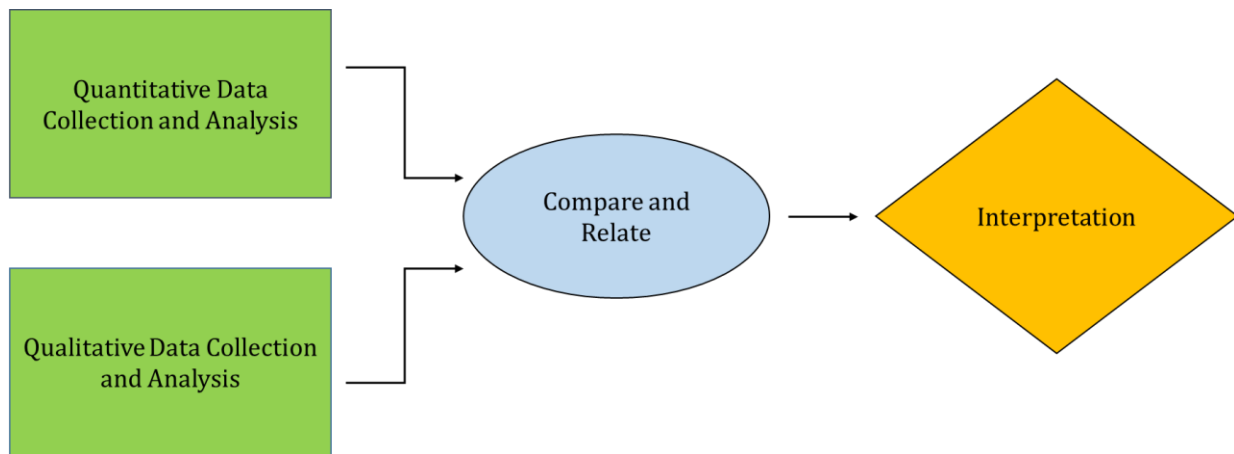


Figure 3.3 Convergent parallel design for mixed methods research

Quantitative and qualitative data were gathered from participants from both the control and experimental groups. Fraenkel and Wallen (2006) noted that experimental designs are an appropriate way of testing the extent to which an independent variable has had an impact on a dependent variable. The use of an experimental design in this study supports the investigation of the possible causal relationships between the type of multimedia stimulus (independent variable) used and test-taker performance and attentional behaviour (dependent variables). In line with between-group design principles, test-takers were randomly assigned to each condition of the independent variable. This limited any potential exposure bias (Howitt & Cramer, 2008). As this study also aimed to observe the role of different categories of response actions (independent variable) in moderating test-taker performance when different multimedia stimuli are used, a within-group design approach was also used. Each participant completed test items with different types of response actions in the item's interaction space (SR, FR, CR). A factorial design was therefore applied¹⁸. Experimental studies that follow a factorial design allow researchers to observe the effect of multiple independent variables on the stated dependent variables. As noted by Coleman (2019), this is a more efficient method of testing hypotheses as two or more things can be studied simultaneously instead of having to conduct multiple separate experiments.

¹⁸ Other types of experiments include: true experimental, pre-experimental, quasi-experimental and single-subject designs (Creswell, 2014; Coleman, 2019).

A factorial design was therefore applied; specifically, a 2x3 mixed factorial design. The between-subjects factor was the multimedia object used in the item stimulus (static vs dynamic). The within-subjects factor was the category of response actions in the item's interaction space (SR vs FR vs CR) (Figure 3.4). Students were randomly assigned to one of the two experimental conditions (static vs dynamic) and then completed all possible categories of response actions (SR vs FR vs CR). Test-taker performance and attentional behaviour acted as the dependent variables. Test-taker scores on an adapted test of scientific literacy were used as a measure of performance for all participants (Study 1A, Study 1B). Eye-movement data were gathered from a subset of the total participant pool to provide various behavioural metrics for those items involved in this experiment (Study 1B). A form of cognitive interview, called a cued-Retrospective Think Aloud, was also conducted with some of these participants (Study 3). This type of cognitive interview will require participants to watch a video of their eye movements and then explain to the researcher what they were thinking at different points of the video. More details on this form of cognitive interview will be outlined in *Section 3.6.3*.

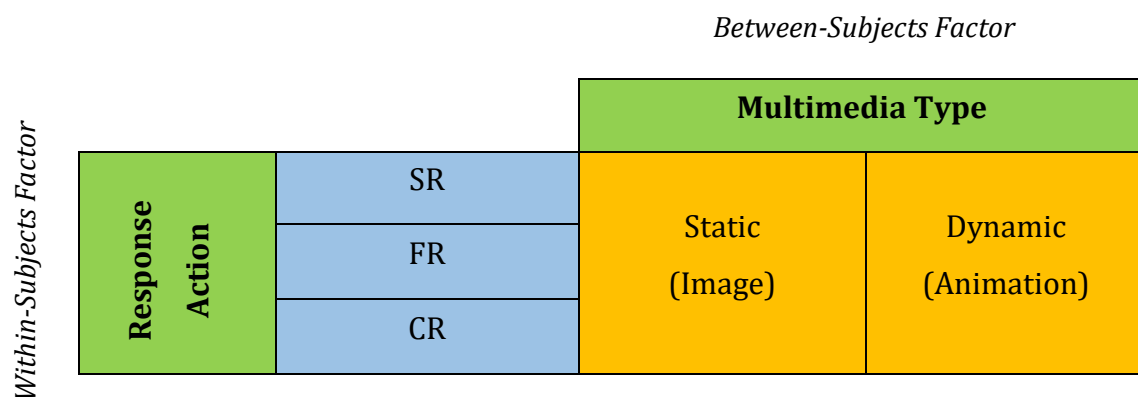


Figure 3.4 2x3 Mixed Factorial Design

Disadvantages of factorial designs can include the complexity of design and interpretation and, at times, the increased number of participants necessary to conduct the study (Coleman, 2019). However, the use of a mixed factorial design with two factors (each with a limited number of levels) alleviated, to some extent, these concerns. Furthermore, the random assignment of individuals to the between-groups condition meant that the majority of threats to internal validity were eliminated (Howitt & Cramer, 2008). However, extra precautions were put in place to minimise all possible threats to internal validity. Internal validity is concerned with the 'question of whether or not the

relationship between two variables is causal' (Howitt & Cramer, 2008, p. 216). Table 3.1 presents the most significant possible threats to internal validity and the ways in which this study controlled for them.

Table 3.1 Controls for threats of internal validity (adapted from Howitt & Cramer, 2008; Creswell, 2014; Coleman, 2019)

Threat	Description	Control
Regression	Extreme scorers revert to the mean of the group.	Random Assignment
Selection	Participants with certain characteristics are selected for certain groups.	Random Assignment
Attrition	Participants drop out during an experiment leaving missing data.	Random Assignment
Order Effects	Response patterns emerge due to the order (e.g. last, first) in which materials are presented.	Materials and tasks will be presented in the same order across conditions
Practice Effects	Any change or improvement in responses resulting from practice or repetition of items or activities.	Materials and tasks will be presented in the same order across conditions
Treatment Diffusion	Participants in different groups communicate with each other which may influence scores.	Communication between participants will not be allowed
Resentful Demoralisation	Participants in the control group underperform as they resent being denied the 'benefits' of the treatment.	The study has no consequences for the participants.

3.4.2 Study 2 and Study 3: Exploratory Eye-Tracking Study

Once the participants involved in Study 1B had completed the TBA, an exploratory eye-tracking study (Study 2) was then conducted using five multi-part simulation tasks. Following that, a short cognitive interview was conducted with a restricted number of participants who were involved in Studies 1B and 2 (Study 3). The eye movement data filled an 'explanatory gap' by providing information on variations in test-taker behaviour

that is not captured by other sources of process data such as computer-generated log files. The collection of data from a think-aloud protocol provided further insights into test-takers' interactions with simulation-type items.

3.5 Research Participants and Sampling

This study used items from an established TBA that measured second-level students' scientific literacy (further details in *Section 3.6*). Consequently, the sample for this study consisted of Irish post-primary students aged between 15 and 16 years of age who attended English-speaking, mainstream schools. As the selection of participants for this study was based on factors other than random chance, non-random sampling techniques were used to recruit participants for all studies (Fraenkel & Wallen, 2006). For Study 1A, participants came from a convenience sample of six second-level schools in Ireland. For logistical reasons, only schools in Leinster and Munster were contacted using the details contained in the Department of Education and Skills' (DES) 2018/19 post-primary database of schools, the most up-to-date database at the time of data collection. Although the use of convenience sampling does not guarantee that the schools contacted contained the student populations that are representative of all those within the Irish education system, such techniques are considered acceptable for social science research (Coleman, 2019). In total, 251 students took part in Study 1A. One school volunteered for their students to be involved in the collection of the eye movement and qualitative data associated with Study 1B ($n=33$), Study 2 ($n=24$) and Study 3 ($n=12$). Figures 3.5 and 3.6 summarise the distribution and contribution of participants across each study.

Purposive sampling techniques were used to recruit potential participants within this convenience sample of schools. Purposive samples involve researchers deliberately studying participants with certain characteristics based on previous research or theory in order to provide the researcher with specific information (Fraenkel & Wallen, 2006). Students in the optional Transition Year (TY) programme were selected for recruitment as this age group are within the appropriate age range for the TBA used and their participation in the current study, which was conducted during school hours, did not interfere with any exam preparation.

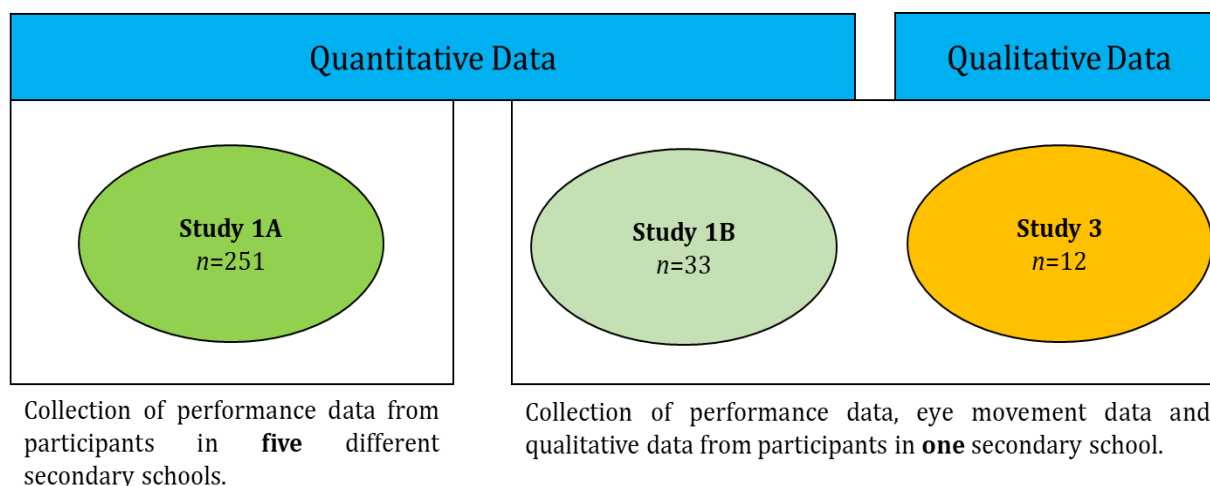


Figure 3.5 Participant numbers and contributions to data collection (Study 1, Study 3)

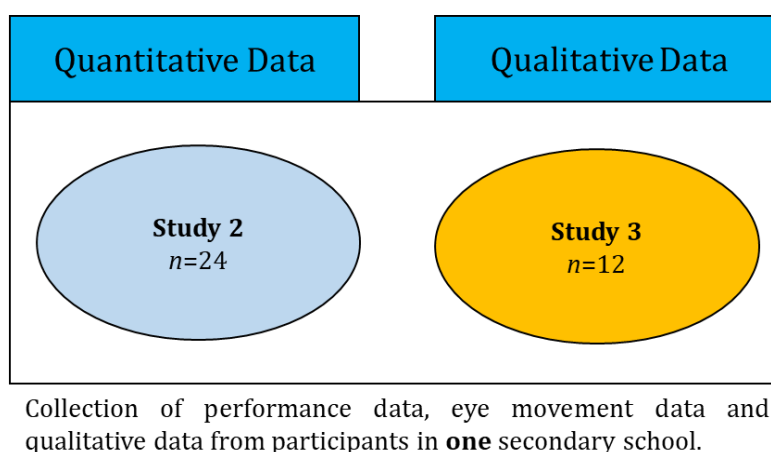


Figure 3.6 Participant numbers and contributions to data collection (Study 2, Study 3)

3.6 Instrumentation

3.6.1 TBA of Scientific Literacy – Study 1 (A, B)

The aim of this study was to explore the use of multimedia objects in test items with varying response actions in TBAs. Publicly available items from the domain of scientific literacy within the Programme for International Student Assessment (PISA) were used to design a TBA suitable for Study 1. As outlined by several researchers including Karakolidis et al. (2021), animations facilitate the measurement of skills and competencies that require test-takers to process sophisticated information. Given the large amount of complex information associated with test items of scientific literacy

(OECD, 2017), items from recent PISA tests of scientific literacy were considered appropriate for use in Study 1.

As with all tests of scientific literacy (Davidsson et al., 2012), test-takers who completed these items were required to understand and apply the relevant information from an item's stimulus to complete certain tasks and questions. Items in tests of scientific literacy aim to capture a range of lower and higher order thinking skills, as in Bloom's Taxonomy¹⁹ (Anderson & Krathwohl, 2001). There is a growing demand for assessments addressing higher-order skills e.g. analysing, evaluating, creating. However, this is often difficult to achieve as the format of these test items can be somewhat limiting. For example, items in scientific literacy tests are primarily in text-form, often providing test-takers with highly complicated passages of text that needed to be fully comprehended before any questions are answered (OECD, 2018; 2016a; 2016b). However, the use of long passages of text may introduce construct-irrelevant variance into the assessment process (Chan & Schmitt, 1997). Therefore, to fully understand the impact of multimedia stimuli on test-taker behaviour and performance, and to determine their true value in test design and construction, the use of animations to replace passages of text requires investigation. Adaptation of PISA items related to the domain of scientific literacy in this study addressed this recommendation and aligned well with the research questions associated with Study 1.

Additional factors influencing the selection of this instrument included the availability of the test items and the population for whom the items were intended (i.e. second-level pupils). Firstly, items from the scientific literacy assessment for the main PISA study in 2015 and for the 2014 field trial²⁰ were available through several OECD (e.g. 2017) publications and their related websites²¹. Like all PISA items, these publicly available items went through a rigorous process to ensure that the content, cognitive demands and contexts of the items were appropriate for 15-year-olds. Psychometric data

¹⁹ Bloom's Taxonomy (1956) is a classification of learning objectives based in the cognitive, affective, and psychomotor domains. In relation to the cognitive domain, the classification system represents a continuum of increasing cognitive complexity: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. Anderson and Krathwohl (2001) revised the taxonomy using verbs to label their categories (rather than the nouns of the original taxonomy) and by switching the order of the two highest thinking skills: Remembering, Understanding, Applying, Analysing, Evaluating, and Creating.

²⁰ The purpose of the PISA Field Trial is to evaluate the appropriateness of the tests, questionnaires and the administrative procedures in each country. This information is then used to refine and improve materials and procedures for the Main Study. The field trials for PISA usually take place 10-12 months before the main study (e.g. Shiel et al., 2016).

²¹ <https://tinyurl.com/2x2wp6kr>

were also available for some of the items used in the main 2015 study. Access to high-quality items and their associated properties was a significant factor in choosing items from PISA to develop the current testing instrument. These items contained many technology-based items, including those that use SR, FR and CR response actions. Moreover, PISA items are designed for use in second-level schools, specifically among students aged 15 to 16-years of age. Bryant (2017) noted that despite being the most tested age group in most education systems, the majority of research on TBAs is not conducted with this population. Using PISA items required the involvement of this under-researched population which also addressed a significant research gap. General details on PISA and its design can be found in Appendix B.

3.6.1.1 Classifying Test Items for a TBA of Scientific Literacy (PISA 2015)

The following items illustrate some of the technology-based items that were asked of students in a single unit (Figure 3.7).

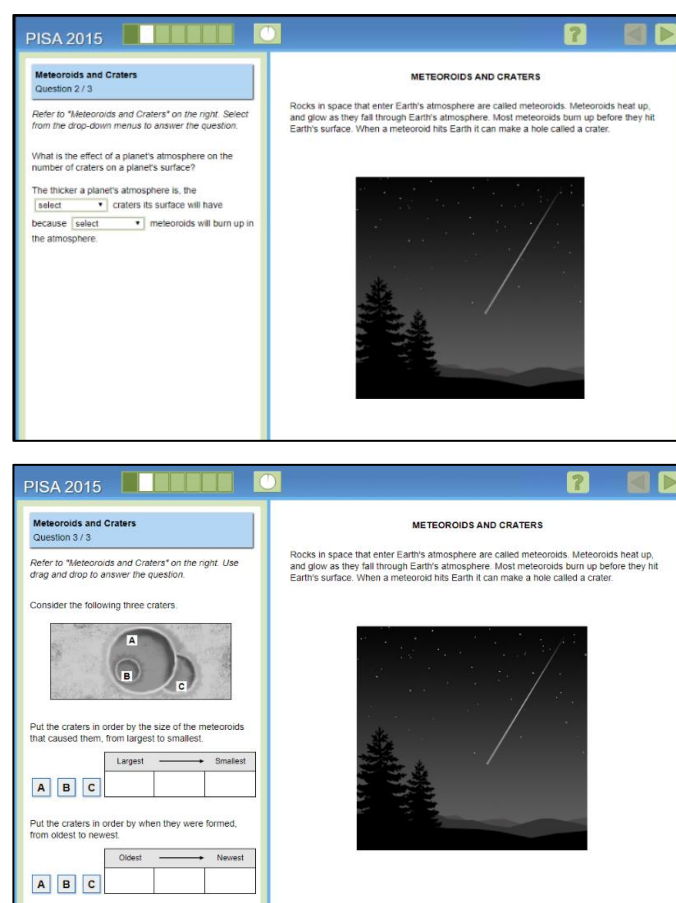


Figure 3.7 'Meteoroids and Craters' (Questions 2-3, PISA 2015 from OECD, 2020)

As shown by Figure 3.7, item stimuli in PISA 2015 were presented on the right-hand side of the computer screen. Item stimuli were usually comprised of a passage of text that provided relevant contextual information and a picture. In other items, these pictures were labelled with text or numbers. Tasks were then contained on the left-hand side of the screen. The OECD (2016b) classified their items according to three broad categories of response formats: simple multiple choice, complex multiple choice, and constructed response. This differs somewhat from the classification framework used in Chapter 2 of this thesis (SR, FR, CR). PISA's definition of 'simple multiple choice' and 'complex multiple choice' do not easily map onto the 'SR' and 'FR' categories used in this thesis. At first glance, 'simple multiple choice' items in the PISA classification system appear synonymous with SR items. However, in the taxonomy of item types used in this thesis, SR items also included items like extended multiple choice (selection of more than one response) and sentence completion (cloze procedures). In contrast, PISA defines these as 'complex multiple choice items' citing examples like:

'responses to a series of related "Yes/No" questions that are scored as a single item... selection of more than one response from a list... completion of a sentence by selecting choices from a drop-down menu to fill multiple blanks... "drag-and-drop" responses, allowing students to move elements on screen to complete a task of matching, ordering or categorising' (OECD, 2016b, p. 54).

With the exception of drag-and-drop items, all of the examples listed above have been classified as SR items in this thesis. This is because the action required by the test-taker is the same i.e. they must 'select' something (Wan & Henly, 2012). Under the PISA framework, these items are considered more complex as they require the selection of *multiple* items. While an increase in frequency does increase the cognitive demand of the item, the *response action* required by the test-taker in the interaction space is still the same. Consequently, these items will continue to be classified as SR items in this thesis as this classification approach is based on relevant literature (e.g. Wan & Henly, 2012; Russell & Moncaleano, 2019). In contrast, drag-and-drop items require test-takers to select something and then *do* something with it e.g. move it into the correct position. As a result of this extra step, drag-and-drop items are considered to be different to SR items and will continue to be classified as FR items.

3.6.1.2 Creating a TBA for Scientific Literacy: Static and Dynamic Stimuli

Nine test units from the 2014 field trial and the 2015 main study were publicly available for viewing from the 2015 PISA cycle (OECD, 2018; 2017; 2016a; 2016b). Each test unit was based on an applied area of scientific knowledge and test-takers had to respond to one or more items associated with this area. Five units were chosen for inclusion in Study 1²² (Table 3.2) taking into consideration the following factors:

- Range of response actions
- Unit content (knowledge, context, skills etc.)
- Feasibility of stimulus modification
- Cognitive demand
- Timing requirements

Table 3.2 Units used in Study 1

Unit (Number of Items)	Knowledge (Context)	Response Actions	Cognitive Demand²³
Bird Migration (1)	Living Systems	1 SR	1 Medium
<i>PRACTICE ITEM</i>	(Global)	1CR	1 High
Meteoroids and Craters (4)	Earth and Space Systems (Global)	2 SR 2 FR	4 Low
Sustainable Fish Farming (4)	Living Systems ²⁴ (Local/National)	1 FR 2 SR	1 Low 1 Medium 1 High
Blue Power Plant (4)	Physical Systems (Local/National)	3 SR 1 CR	1 Low 3 Medium
Groundwater Extraction and Earthquakes (4)	Earth and Space Systems (Local/National)	2 SR 1 FR 1 CR	1 Low 3 Medium

²² In the main 2015 PISA study, students completed 8-10 units (approximately 30 items) resulting in 'about one hour of testing' (OECD, 2016b, p. 57). Five units were included in this study instead which resulted in a 30-minute test (18 items). This was less than a standard class period in Ireland (40-50 minutes), thus minimising any disruption to the students' school day as a result of their participation in this study.

²³ The cognitive demand of items is influenced by four factors according to the OECD (2016b): (i) the number and degree of complexity of the elements of knowledge in the item, (ii) students' level of familiarity with the knowledge involved (iii) the cognitive operation required by the item, e.g. recall, analysis (iv) the extent to which forming a response depends on models or abstract scientific ideas.

²⁴ The final question in this unit is classified as a 'Physical Systems' category.

The units summarised in Table 3.2 included items that had a range of response actions available and represented all aspects of the previously outlined scientific literacy framework. The researcher also had the necessary content knowledge for each of these units to facilitate an informed decision-making process regarding the required stimulus modifications for this study. The ‘Bird Migration’ unit contained one item which was used as a practice item and did not contribute to the participants’ scores on the TBA. No modifications were made in relation to the response actions associated with the items contained in each unit for Study 1. These were replicated in the testing platform used (see Section 3.6.1.3 for further details). Each of the units listed in Table 3.2 had between one and three standard text-image stimuli²⁵ providing contextual information for different items within the unit. However, to develop dynamic stimuli (animations) for each of the units, the researcher worked with an Irish-based animation company²⁶.

Development of the dynamic stimuli involved the researcher and the animation company working together to ‘storyboard’ each of the animations based on the original image and text used in the original PISA unit. The images in the original PISA units were the first ‘scene’ for most of the animations. Each animation aimed to display a dynamic representation of the contextual information contained in the accompanying voice-over which read the exact text found in the original PISA item²⁷. The animations offered a visual representation of key concepts and ideas as they were read out (e.g. displaying a tally card in the ‘Bird Migration’ animation when the voiceover stated that ‘... *sightings of tagged birds together with volunteers’ counts...*’). To ensure the quality and accuracy of the animations, the researcher and three different people with expertise in educational research and/or post-primary science content reviewed each of the animations²⁸. To ensure that there was comparability between the conditions, still images from the designed animations were used in the static version of the test. The still image used was comparable to the image used in the original PISA 2015 version. Figure 3.8 shows the

²⁵ One unit, ‘Blue Power Plant’, had a short 2 second gif demonstrating the process of salt and freshwater osmosis. Students had to click a magnifying glass to see this dynamic stimulus which represented one of the many pieces of contextual information presented to students.

²⁶ baboom.ie; <http://baboom.ie/>.

²⁷ Minor adjustments to the text were made to reflect the change in stimulus e.g. ‘This image shows...’ became ‘This animation shows...’.

²⁸ These people were: the doctoral student, the two supervisors of the project, and an independent researcher with previous teaching experience of post-primary science courses in Ireland.

original PISA item stimulus, the stimulus used in the static condition of the current study and a still image from the dynamic condition of the current study.

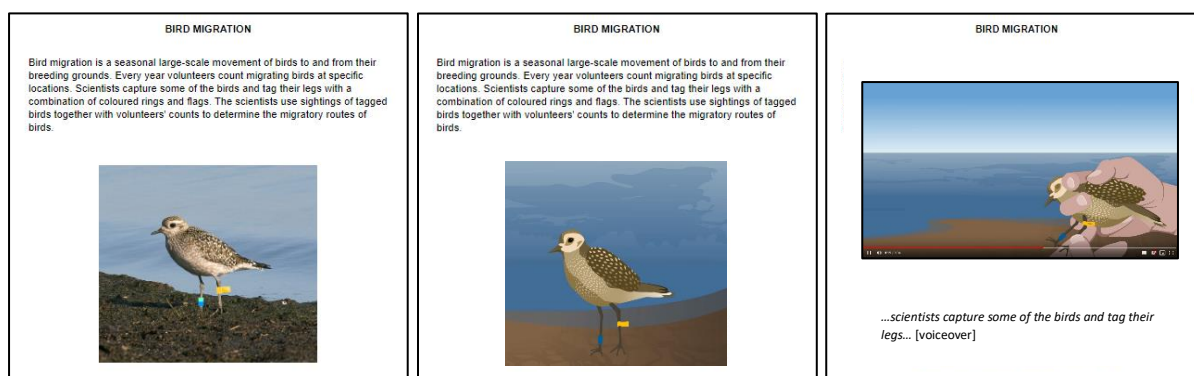


Figure 3.8 ‘Bird Migration’ – Original PISA 2015 stimulus, stimulus for static condition, stimulus for animated condition

3.6.1.3 Testing Platform

For Study 1, the units used for the TBA of scientific literacy were presented in the following order to the participants regardless of the condition they are in:

- Bird Migration (Practice Item)
- Meteoroids and Craters
- Sustainable Fish Farming
- Blue Power Plant
- Groundwater Extraction and Earthquakes

Key requirements for creating an assessment environment that would be engaging and would support the experimental comparisons being investigated were identified at the start of the study (informed by Karakolidis et al., 2021). They were:

- The platform should be able to support the high definition animated videos and the audio files that would accompany the practice statements.
- Test-takers should be able to navigate between items within a unit but not between units (as per PISA guidelines).
- The videos and the statements should be easily accessible by the test-takers with minimum scrolling, as recommended by the relevant research literature (Bridgeman et al., 2003; Sanchez & Goolsbee, 2010).

- The platform should adhere to responsive web design principles²⁹.
- Participants should be randomly assigned to each condition by the system.

Off-the-shelf versions of commercial platforms, such as *eSurveyCreator* and *SurveyMonkey*, did not meet the necessary requirements. As a result, another software company (*Psycholatte*) was approached on the basis of previous work they had conducted with a research centre in Dublin City University. Figure 3.9 contains screenshots of the platform used for both conditions³⁰. See Appendix C for screenshots of all items across both conditions. Appendix D contains screenshots of the introductory screens of the TBA.

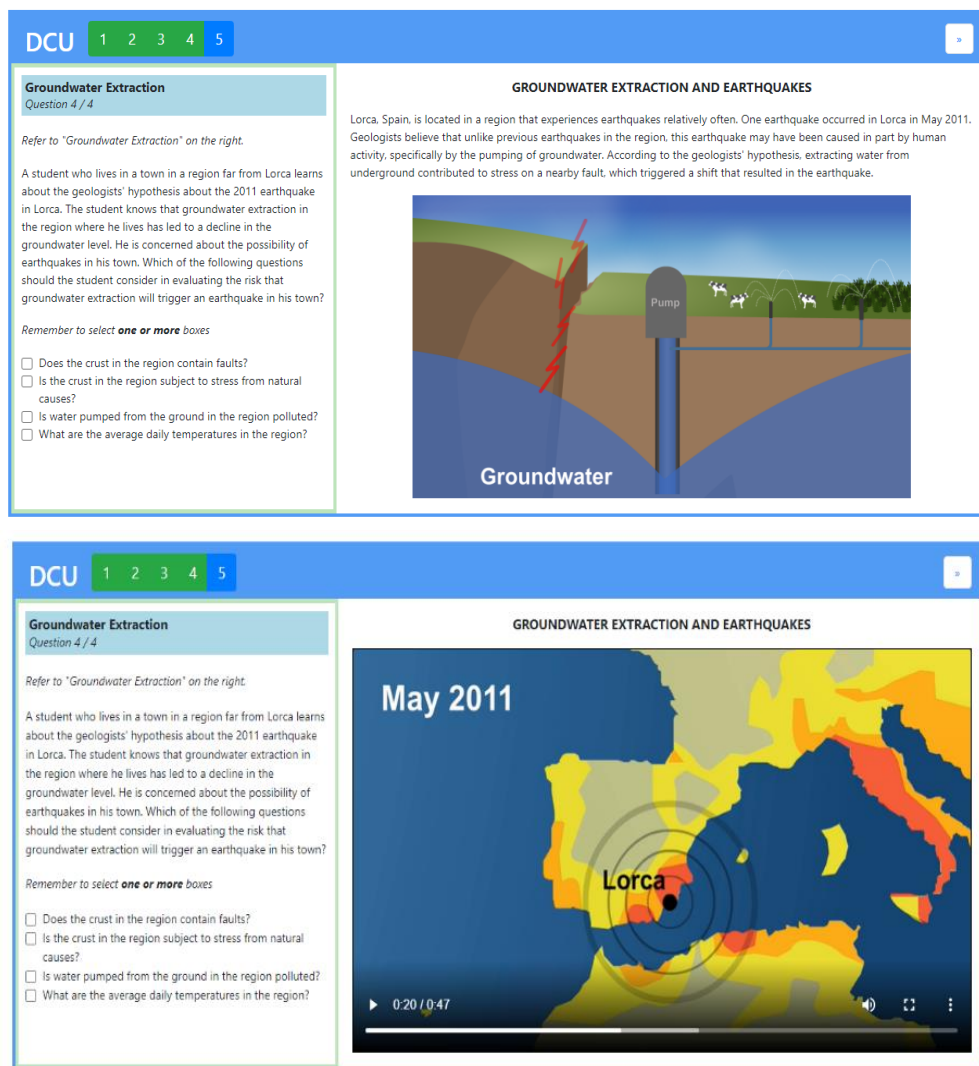


Figure 3.9 Item from ‘Groundwater Extraction and Earthquakes’ unit (Static, Dynamic)

²⁹ Responsive web design ensures that online environments are modified to reflect a device’s screen size, platform and orientation (Gregory, 2019).

³⁰ This is an adaptation of an original work by the OECD. The opinions expressed and arguments employed in this adaptation are the sole responsibility of the authors of the adaptation and should not be reported as representing the official views of the OECD or of its member countries.

3.6.2 Simulation-Type Items – Study 2

Study 2 aimed to investigate test-taker engagement with simulation-type items. PISA 2015 developed a number of simulation-type items for use in their TBAs of Scientific Literacy. The ‘Running in Hot Weather’ simulation unit was chosen as the stimulus for Study 2 as there were more items available for public use than the other unit available online (e.g. ‘Slope Face Investigation’). An example simulation type item can be seen in Figure 3.10. Appendix B contains an overview of how simulation-type items addressed the competencies, knowledge and skills included in PISA 2015’s Scientific Literacy Framework.

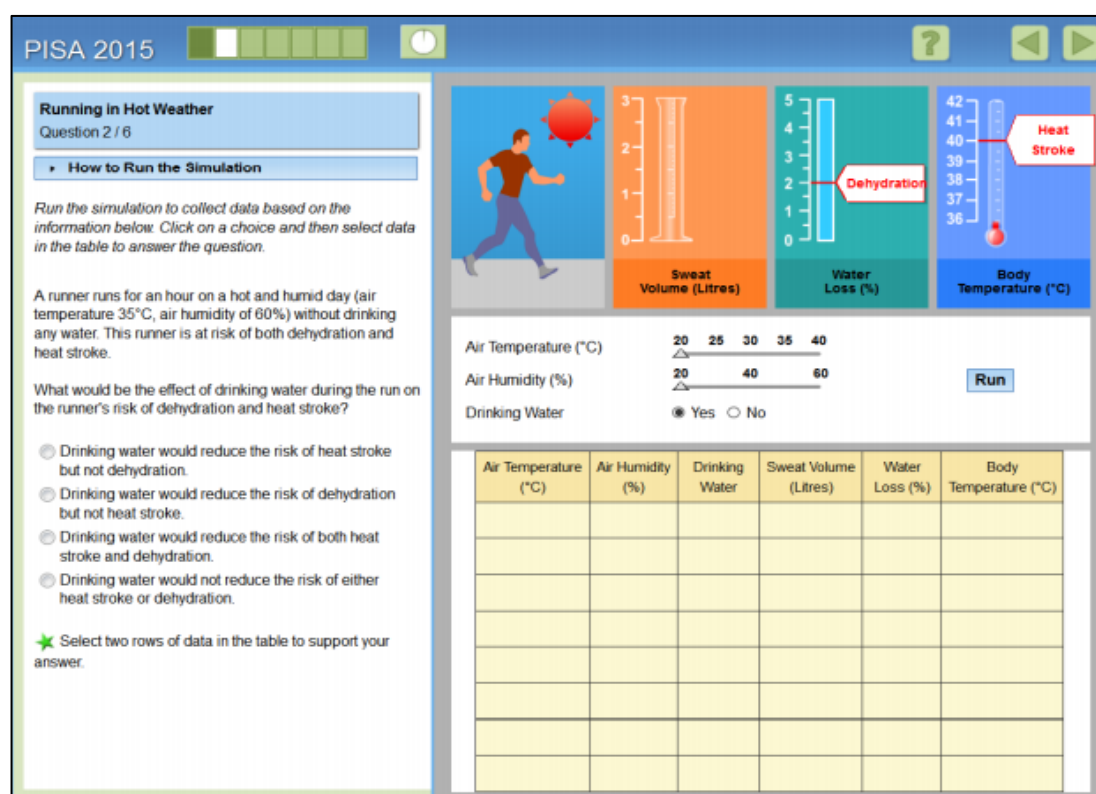


Figure 3.10 ‘Running in Hot Weather’ (Question 2, PISA 2015 from OECD, 2020)

Five tasks were contained in this unit for Study 2 (see Appendix E for screenshots of all items). As it was not possible for these items to be downloaded or modified, participants accessed the unit directly through the PISA website³¹. Before beginning the unit, students were introduced to the simulation controls and asked to practice setting each control. Help messages were displayed if students did not perform the requested

³¹ <https://tinyurl.com/2x2wp6kr>

actions within 1 minute. After completing the tasks required for each item, a brief message indicating the participants' performance was displayed along with a brief explanation of the correct response to the task.

3.6.2.1 Classifying Simulation-Type Items

It must be noted that PISA 2015 classified their simulation type items in a way that is somewhat inconsistent with research literature. In the question represented in Figure 3.10, students are asked to run the simulation holding the air temperature and humidity constant using specified values. They must also manipulate the variable of whether the runner drinks water to determine what the difference in running experience will be. The simulation shows that running under the specified conditions without drinking water leads to both dehydration and heat stroke. In contrast, drinking water reduces the risk of dehydration but not the risk of heat stroke. Students must run the simulation twice in order to collect the data needed to answer the multiple choice question on the left-hand side of the screen. Despite labelling it a simulation throughout their literature, PISA 2015 classifies this item format as a combination of a 'simple multiple choice' and 'open response'.

While the item may appear to look like an SR type item due to the presence of a multiple choice question on the left-hand side of the screen, test-takers are required to do more than just 'select' an answer to complete the task. In fact, it requires 'specialised interactions for response' (Measure Progress/ETS Collaborative, 2012, p. 9). More specifically this item fully aligns with Levy's (2012) definition of simulation-type assessments³². Firstly, this item is highly interactive, requiring test-takers to engage in multiple different response actions (dragging sliders, selecting variables etc.,). This should prompt a reply from the system that will influence their next response (what variables to select next, selecting the appropriate answer from the list of four options, highlighting with a green star what rows of data support their selection). In doing so, the test-taker is thought to be engaging one of the key competencies of the science literacy framework (*Interpret Data and Evidence Scientifically*) in an authentic manner that replicates a real world scenario.

³² Levy (2012) defines simulation based assessments as one involving static or dynamic stimuli that allows the test-taker to be 'presented with, work with, or produce a work product that contains a simulation of a real world scenario' (p. 10).

3.6.3 Cognitive Interview – Study 3

Study 1B and Study 2 involved the collection of eye-movement data (see Section 3.8.2 for further details). Eye movement patterns and fixations can be used to gain an understanding about how respondents complete a particular task or activity using objective numerical data. However, it is recommended that eye-movement data be combined with other data to aid in its interpretation (Elbabour et al., 2017). Qualitative data from a think-aloud protocol were collected to achieve this in Study 3. Think-aloud protocols are most commonly used when attempting to detect usability problems in web design (Nielson, 2012). Think-aloud protocols have been traditionally classified into two types: concurrent think-aloud (CTA) and retrospective think-aloud (RTA). In CTAs, participants are asked to verbalise their thoughts *while* they are doing tasks. In RTAs, the participants provide a description of their experiences doing the tasks *after* each or all of the tasks are completed (Elbabour et al., 2017). Olsen et al. (2010) note that both approaches are relatively simple methods of gaining insight into the participants' thought processes regarding task completion. However, each of these methods offers its own set of problems or limitations which should be considered when selecting a methodology. For example, it is important to remember that cognitive processes are quicker than verbal processes (Sternberg, 2009). As a result, participants might be thinking about more than they are able to verbally express in CTAs. The act of trying to verbalise thoughts may also interfere with task performance (Olsen et al., 2010). In RTAs, the participant is required to remember their experiences rather than communicate their moment-to-moment decisions and actions as they happen. This means that important information may be forgotten or misremembered (Elbabour et al., 2017).

Olsen et al. (2010) recommend using RTA protocols in eye-tracking studies as CTAs often result in participants producing confounding eye movements like 'looking away from the screen to describe something to the researcher or by focusing on certain areas of the screen while describing their thought processes regarding that area' (p. 46). Therefore, the retrospective think-aloud method was used in this study to aid in the analysis and interpretation of objective eye movement data. A specific type of retrospective think-aloud was deployed in Study 3 to address the previously mentioned shortcomings of this method and to fully exploit the research opportunities available with eye-tracking equipment. In order to aid participant memory, cued RTA (c-RTA) methods have become more common in research (Olsen et al., 2010). In a c-RTA, the participant is

presented 'with a form of replay of the interactions they previously performed in order to help cue their memory' (Olsen et al., 2010, p. 46). The participants are asked to respond to this replay, stating out loud to the researcher what they were thinking at different points of the video. The replay (or 'cue') can take many forms in a c-RTA. Using a true experimental approach ($n=24$), Olsen et al. (2010) compared four types of c-RTA methods (no cue, video-cued RTA [a video replaying their actions], gaze-plot cued RTA [eye movement on still image], and gaze-cued RTA [superimposed eye movements on a screen video]) in a usability study. The authors found that the gaze-cued RTA method was more effective than the alternatives. This particular 'cue' helped participants to verbalise almost double the number of words compared to those participants in the other conditions and it was also very effective in helping the participants identify usability problems. A later study by Elbabour et al. (2017) supported the value of a gaze-cued RTA.

As a result, a gaze based c-RTA was considered the most appropriate think-aloud protocol to use in this study. Participants were asked to watch a video of their test taking actions for 4 items (one SR, FR, CR and simulation-type item) with their eye-movements superimposed onto the video. While watching the video, the researcher asked the participants to recall what they were thinking at each point in the video.

3.7 Equipment

An eye-tracker was required to collect the eye movement data needed for Study 1B and Study 2. There are a wide variety of commercial eye trackers available to researchers (Carter & Luke, 2020). Trackers mainly vary in their speed of data acquisition, as measured in Hertz (Hz), and their set-up e.g. stationary eye trackers, mobile eye trackers. Taking into consideration the funding available, the intended participants of the current study and the data required, a stationary eye tracker was considered to be the most appropriate for the current study. The eye tracker used for the current study was the *Tobii Pro Fusion*, a screen based eye tracker that tracks both eyes while tolerating a variety of head movements and a wide range of physiological variations e.g. eye colour, use of bi-focal glasses, use of contact lenses (see Figure 3.11). Accuracy and precision test reports on this eye-tracker demonstrated that it can collect highly

accurate (within $.3^\circ$) and precise (within $.2^\circ$)³³ eye movement data (Tobii AB, 2020). Its sampling rate of 120Hz allows for 120 data points for each eye to be captured every second, allowing for a more accurate estimate the true path of the eye when it moves. Although eye trackers with higher sampling rates were available, a lower sampling rate was deemed acceptable as the study was mainly interested in recording where a participant looked, thus negating the need for accuracy beyond milliseconds. All eye movement data were recorded in the *Tobii Pro Lab* software.



Figure 3.11 Tobii Pro Fusion Eye-Tracker

3.8 Measures and Variables

For Study 1, two independent variables³⁴ were considered. The between-subjects factor for this experiment was the multimedia object used in an item's stimulus. This independent variable had two categories: static (images, text) and dynamic (animations). The second independent variable was a within-subjects factor related to the type of response action permitted in the item's interaction space. It had three categories: SR, FR and CR. The dependent variables³⁵ for Study 1A were test-taker performance on the TBA and for Study 1B test-taker performance was paired with test-taker attentional behaviour during this test. Participant overall scores on the TBA of scientific literacy were used as an outcome measure for test-taker performance. Eye movement data were used as the

³³ Please see Appendix G for more details on the importance of accuracy and precision when using eye-tracking equipment.

³⁴ Fraenkel and Wallen (2006, G-4) explain that an independent variable is any 'variable that affects (or is presumed to affect) the dependent variable under study and is included in the research design so that its effect can be determined'. It can also be called the experimental or treatment variable.

³⁵ A dependent variable refers to any variable that that is 'affected by the independent variable' (Fraenkel & Wallen, 2006, G-2). It may also be referred to in literature as the criterion or outcome variable.

measure of test-taker attentional behaviour. For Study 2, participants completed five tasks with simulation-type items. Performance data and eye movement data for each item in these tasks were collected. Study 3 involved the collection of qualitative data using a form of cognitive think-aloud from participants who were involved in Study 1B and Study 2. These outcomes measures will now be discussed in more detail.

3.8.1 Performance Data

3.8.1.1 Study 1 (A, B) – Scores on a TBA of Scientific Literacy

Participants completed 16 items in the TBA used for Study 1A and Study 1B. One item (from the ‘Bird Migration’ unit) was used as a practice item to allow the participants to become more familiar with the testing platform. Therefore, the maximum possible score that participants could achieve on the TBA used in Study 1A and Study 1B was 15. Each correct SR and FR item was identified by the testing platform and given a score of 1. Responses to CR items were marked as incorrect or correct by the researcher according to the PISA 2015 guidelines for those items (OECD, 2017). For example, when completing item 4 of the ‘Blue Power Plant’ unit, students are asked to explain why the power plant shown is considered to be more environmentally friendly than power plants that use fossil fuels. According to the guidelines for this item, students must provide an explanation that identifies a way in which plants that burn fossil fuel are more harmful to the environment than the new power plant illustrated e.g. fossil fuels release carbon dioxide/ greenhouse gases that can harm the environment. Alternatively, the student can identify a feature of the new power plant that makes it more environmentally friendly e.g. the plant runs on a renewable form of energy. The application of these scoring guidelines to a randomly selected sample of responses was reviewed by the researcher and two supervisors.

3.8.1.2 Study 2 – Scoring Simulation-Type Items

Five tasks requiring the use of simulations were included in Study 2 (‘Running in Hot Weather’). Each task had an item that the test-takers had to complete. In most cases, it was a multi-part item. All tasks required test-takers to complete an SR-type item part (Part ‘A’). For Tasks 2-5, test-takers also had to select simulation data to support their answer for Part A (Part ‘B’). Tasks 3-5 had a CR part to the item (Part ‘C’), where participants had to explain the scientific topic or principle underlying Parts A and B. In

PISA 2015, a *single* score was given for test-taker performance on Parts A, B and C using a partial credit scoring system, where the maximum score was 2 (OECD, 2017; 2015).

A modified approach to scoring the items used in Study 2 was applied to ensure that item parts with different response actions were scored separately. In Part A and Part B, the participant had to complete a multiple-choice question by *selecting* their answer and then *select* the line(s) of data that informed their choice. For Part C, the participants had to *type* their justification or explain the underlying scientific process. While the item associated with each simulation task addressed one key competency in PISA 2015 (see Appendix B), using a single scoring system for an item that contained different response actions was considered problematic. Literature indicates that the efficacy of different response actions in measuring test-taker proficiency is still unknown (e.g. Wan & Henly, 2012). To minimise the impact of any confounding variables on test-taker scores, test-taker responses to Part A and Part B were considered separately to Part C. A partial credit scoring system (ranging from 0 to 2) was in place for Parts A and B (Part AB), based on the guidelines from PISA 2015 (OCED, 2017; 2015). Part C of the item was marked simply as ‘correct’ or ‘incorrect’ based on the PISA 2015 scoring guidelines.

3.8.2 Process Data

Eye tracking is an experimental method of recording eye motion and gaze location across a particular task (Carter & Luke, 2020). For Study 1B and Study 2, eye movement data were used to gather the process data needed to better understand secondary school students’ subconscious mental processes when engaging in assessment activities involving different item types and multimedia stimuli. Eye tracking research has experienced a surge in the past decade as the equipment and associated software has become cheaper and more user-friendly. Yet, Orquin and Hölmqvist (2018, p. 1645) have cautioned that this ‘surge in eye-tracking research has not, however, been equalled by a growth in methodological awareness’. The authors have argued that many eye tracking studies employ practices that pose a significant threat to their validity. In particular, the large amount of data generated in eye tracking studies can be overwhelming and a lack of a clear data extraction and analysis plan can increase the risk of Type I errors (Carter & Luke, 2020). To minimise such risks to the validity of this piece of eye movement research, a simplified process of preregistration was applied in this research study, in accordance with Carter and Luke’s (2020) guidelines. Preregistration requires

researchers to publicly define their research questions and analysis plans before observing outcomes (Kryptos et al., 2019). For eye tracking studies, effective preregistration would involve the identification of the key events and areas of interest prior to data collection. The variables that will be used for data analysis should also be selected before data collection, along with a justification for each variable chosen (Carter & Luke, 2020). The times and areas of interest for this research will now be outlined along with the selected eye tracking metrics. Further details on the other key issues associated with the current research's use of eye movement data (e.g. ensuring data quality) can be found in Appendix F.

3.8.2.1 Times of Interest (TOIs)

Each participant generated a recording of their eye movements (which can also be called a 'timeline'). A Time of Interest (TOI) is created by specifying start and end events within a recording or timeline. While a common approach in eye tracking studies is to 'fix' the exposure time or TOI (so that the participant only sees the stimulus of interest for a predetermined period of time), in accordance with Orquin and Hölmqvist's (2018) recommendations, free exposure time was used. While this demanded some extra data management steps to ensure comparability (as the duration of the TOIs differed between participants), it is more appropriate for use in behaviour-orientated studies. For Study 1B, TOIs were calculated for each *item* that the participant completed. As a result, 16 TOIs were created for each participant in this study.

Identifying the TOIs for the simulation-type items used for Study 2 *a priori* was more challenging, particularly as there were no other similar studies conducted at the time of data collection to act as a guide. While creating a TOI for each simulation that an individual participant runs would have been the most straightforward way to subdivide the events involved in this item, this would be somewhat problematic for data analysis as it was impossible to predict in advance how many simulations an individual would actually do. Unlike the items in Study 1B, there was a much greater range of participant behaviours possible within Study 2's items. Such a range of behaviours could introduce significant noise into the eye movement data collected unless some efforts were undertaken to subdivide the item into discrete TOIs that were clearly justified and associated with specific eye movement metrics (Orquin & Hölmqvist, 2018).

To better account for and understand the variations in participant behaviour that could have emerged with these items, two TOIs were created for each of the five tasks in this unit. The first TOI for each task began when the item on the left hand side of the screen was first presented and ended when the first simulation was executed ('Orientation'). The second TOI began when the results of the first simulation were presented to the time that the participant navigated away from the task ('Output'). The first TOI for each task was considered to act as an orientation time, where the participant became familiar with the investigation required of them and to then 'set up' their first simulation. The duration of this TOI, as well as what was attended to, would provide insight into what test-taking strategies and behaviours were employed by the participant. Of particular interest within this TOI was whether any of the participants' eye movements aligned with the construct that was alleged to be measured by the item. For example, if the item was designed to assess participants' skills in designing scientific enquiries, then participants' subconscious attendance on areas of the simulation relevant to a hypothesis needed to answer the question was of interest during this TOI.

The second TOI contained all the behaviours and events associated with the participant's efforts to answer the task's item based on the output of the first simulation. The second TOI would therefore include any subsequent simulations that the participant may have conducted on the basis of the output from the first simulation along with their final answer to the target question(s). Although the events within this TOI could also vary significantly between participants (e.g. number of simulations), the TOI aimed to illustrate the focus of the participants' attentions for the duration of the investigation and whether or not they demonstrated their competencies in the skills these items were designed to assess. For example, in Item 2 of this unit (See Appendix C), was designed to assess if participants can 'interpret data and evidence scientifically' (OECD, 2015). This skill can only occur after the first simulation has been executed as the participant attempts to find the information they need to answer the target question. Similarly, Item 3 in this unit assessed if students could 'evaluate and design scientific enquiry' (OECD, 2016a). Again, this competency can only be assessed *after* the simulation provides some data to guide future simulations e.g. changing the air temperature. Depending on the competencies that each item was designed to assess, the metrics used to summarise the eye movement data in the second TOI of each item did vary. However, this ensured that that any eye movement data collected could be used to examine whether or not the

respondents engaged as expected in order to assess the competencies associated with that item.

3.8.2.2 Areas of Interest (AOIs)

Areas of Interest (AOIs) are used to link eye movement data to specific parts of the presented stimulus (e.g. the time spent looking at a particular location on screen) (Hessels et al., 2016). Most eye tracking software, including *Tobii Pro Lab*, allows the user to predefine these AOIs. After data collection, the software processes the eye movement data to provide a description of how each participant interacted with these areas. Hessels et al. (2016) and Carter and Luke (2020) asserted that AOIs should be carefully demarcated in eye tracking studies prior to data collection. If they are not, the outcome of a research study could be compromised. As a result, the current study paid particular attention to the size and location of the AOIs used in Study 1B and 2. These will be described after a brief summary of how relevant guidelines on AOI construction for this study were decided upon.

Hessels et al. (2016) noted that defining AOIs is a significant problem in eye tracking research. Choices on AOI shape, size and placement can vary even across experts even when identical stimuli are used. In relation to size, issues of selectivity and sensitivity must be considered and Orquin et al.'s (2016) work offers important insight on this subject. Smaller AOIs (relative to the size of the overall presentation area) increase selectivity as they are more specific but they risk losing valid eye fixations. Larger AOIs increase sensitivity but may include extraneous data (Hessels et al., 2016; Orquin et al., 2016). However, Orquin and Hölmqvist (2018) believe that larger AOIs are more appropriate to use given the capabilities of current eye tracking machines. Similarly, if multiple AOIs are usually present on a stimulus page, the spaces *between* AOIs also needs to be accounted for and related to the accuracy of the eye tracking software. Orquin and Hölmqvist (2018, p. 1653) explained that AOIs with narrow margins around the location of interest e.g. less than $.5^\circ$ (less than 20 pixels approximately) between what is and is not considered to be of interest, may cause the eye tracker to 'fail to detect fixations falling just outside the object'. This would lead to eye movement data full of false negatives. Similarly, if AOIs are placed too close together, false positives may occur where fixations are incorrectly assigned to neighbouring AOIs.

To address these issues, a systematic approach to the sizing and placement of AOIs was employed. Taking into consideration the manufacturer's guidelines for the *Tobii Pro Fusion* machine (Tobii AB, 2020) and the experiences of other researchers (e.g. Orquin & Hölmqvist, 2018), borders with a minimum width ranging between .50° and 1° (approximately 20-50 pixels on screen, depending on the distance between the participant and the screen) were included in and between AOIs drawn around target areas of the stimulus. This also compares with the calibration cut-off figure used i.e. eye movement data should be relatively accurate within .5°. While the *Tobii Pro Lab* software permitted 'hand drawn' AOIs, the target areas in the stimulus materials were (for the most part) contained within quadrilateral shapes e.g. answer boxes, grids. Therefore, the AOIs used in Study 1B and 2 were all quadrilaterals. *Tobii Pro Lab* treated each AOI created within each item as separate from all others. For the purposes of the current study however, AOIs that addressed similar issues within an item were aggregated under a 'tag' to facilitate comparative analyses across items and participants. A further explanation of how these construction criteria and tags were applied to the stimulus materials for these studies will now be provided.

Study 1B

Each of the 15 items in Study 1B had a similar format, which was replicated from PISA 2015's original items. The right hand side contained the stimulus materials that provided context to the overall unit or specific question. This varied according to the condition (Dynamic vs Static). The left hand side contained the item number along with a variation of instructions, question stems (in text or image form) and interaction spaces (containing SR, FR or CR response actions). At the top of every item, there was a navigation bar that included a progress bar and a 'Next' button. Figure 3.12 demonstrates the AOI placement for an SR item across both conditions³⁶.

³⁶ This is an adaptation of an original work by the OECD. The opinions expressed and arguments employed in this adaptation are the sole responsibility of the authors of the adaptation and should not be reported as representing the official views of the OECD or of its member countries.

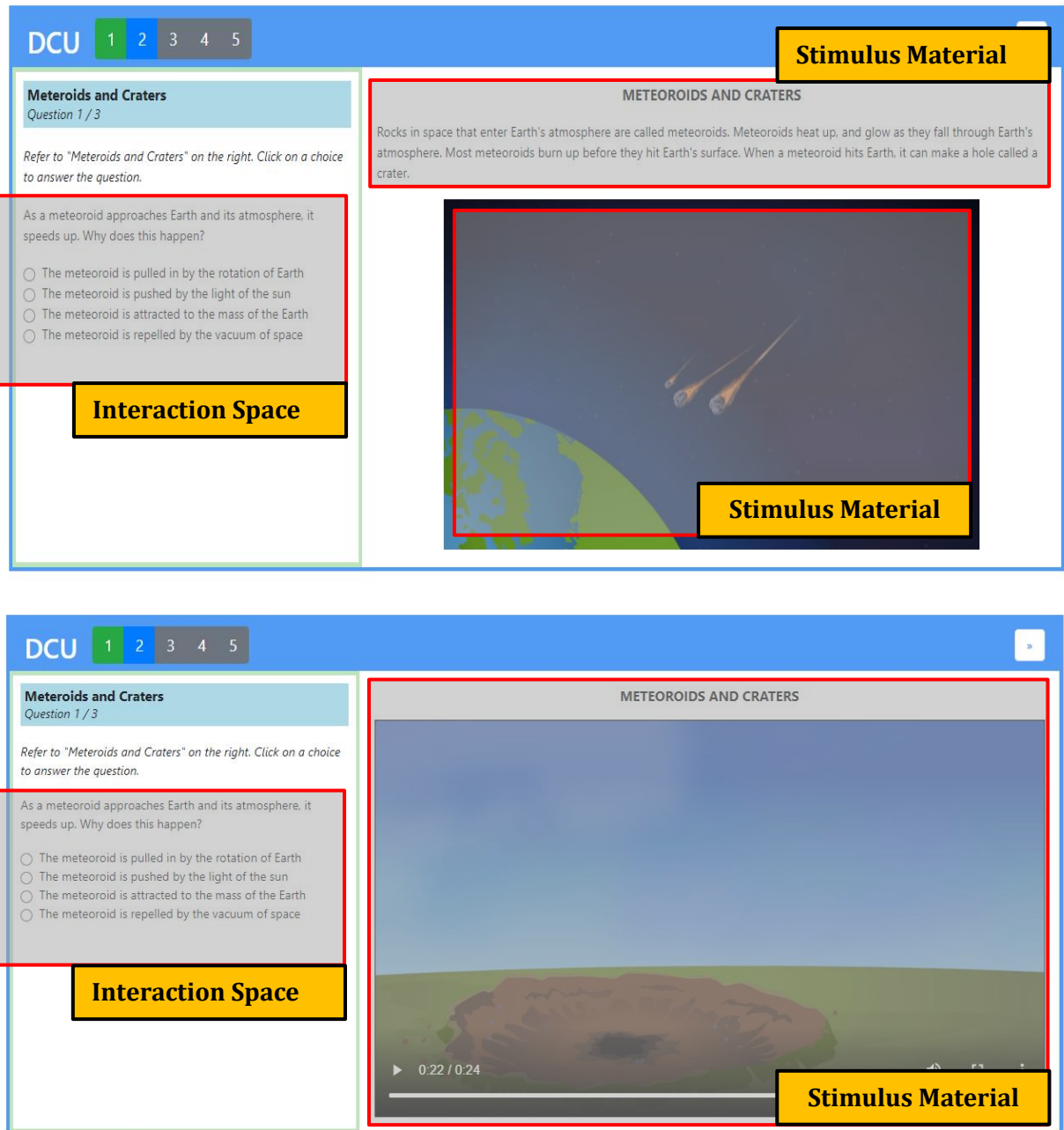


Figure 3.12 Placement of AOIs for an SR item in Study 1B (Static, Dynamic)

As seen in Figure 3.12, AOIs were drawn around the *Stimulus Materials* in each item. In the Dynamic condition, this included the video and its title. In the Static condition, separate AOIs were created for the presented text and image stimuli. For certain items, the size of the AOIs for the images was slightly smaller than the image to ensure that there was sufficient space between the text and image AOIs. The visual information within the picture not captured by the AOI had no new or relevant content. AOIs were also constructed around the *Interaction Space*. AOIs for SR items captured the entire range of

possible response options. For MCQs, it was not possible to separate out the individual response options as they were too close together (e.g. Figure 3.12). For other SR items involving drop-down menus to ‘fill in the blanks’, the AOI covered the entire sentence as it would have been conceptually inappropriate to separate these out. For FR items, the AOIs were relatively large but confined to the area that permitted the movement of any portable objects. For CR items, the AOI covered the text input box. To avoid any ‘false positives’, a minimum border of 20 pixels between all AOIs was present. AOIs that contained any text, such as the AOIs surrounding question stems, had a border between 20 and 40 pixels in width to ensure that the eye tracker did not miss any relevant fixations just beyond the edge of the text (thus avoiding ‘false negatives’).

Study 2

Study 2 was accessed directly from the PISA website (OECD, 2020) and asked participants to complete the ‘Running in Hot Weather’ unit. The AOIs for this item were static as their position on the screen did not change as the participant executed their simulations. Figure 3.13 shows the size and location of the AOIs for items in this study during the ‘Orientation’ phase described in Section 3.8.2.1. On the left hand side of the item, the AOI covered the details participants needed to run the simulation and answer the question asked of them. On the right-hand side of the screen, the simulation controls were another AOI. The areas where the test-takers would find the data necessary to answer the test item were represented by ‘Relevant’ AOIs. By highlighting these two areas on the screen, it was hoped that a better understanding of how participants use the information presented to prepare, design and execute a scientific inquiry as per the aims of this unit.

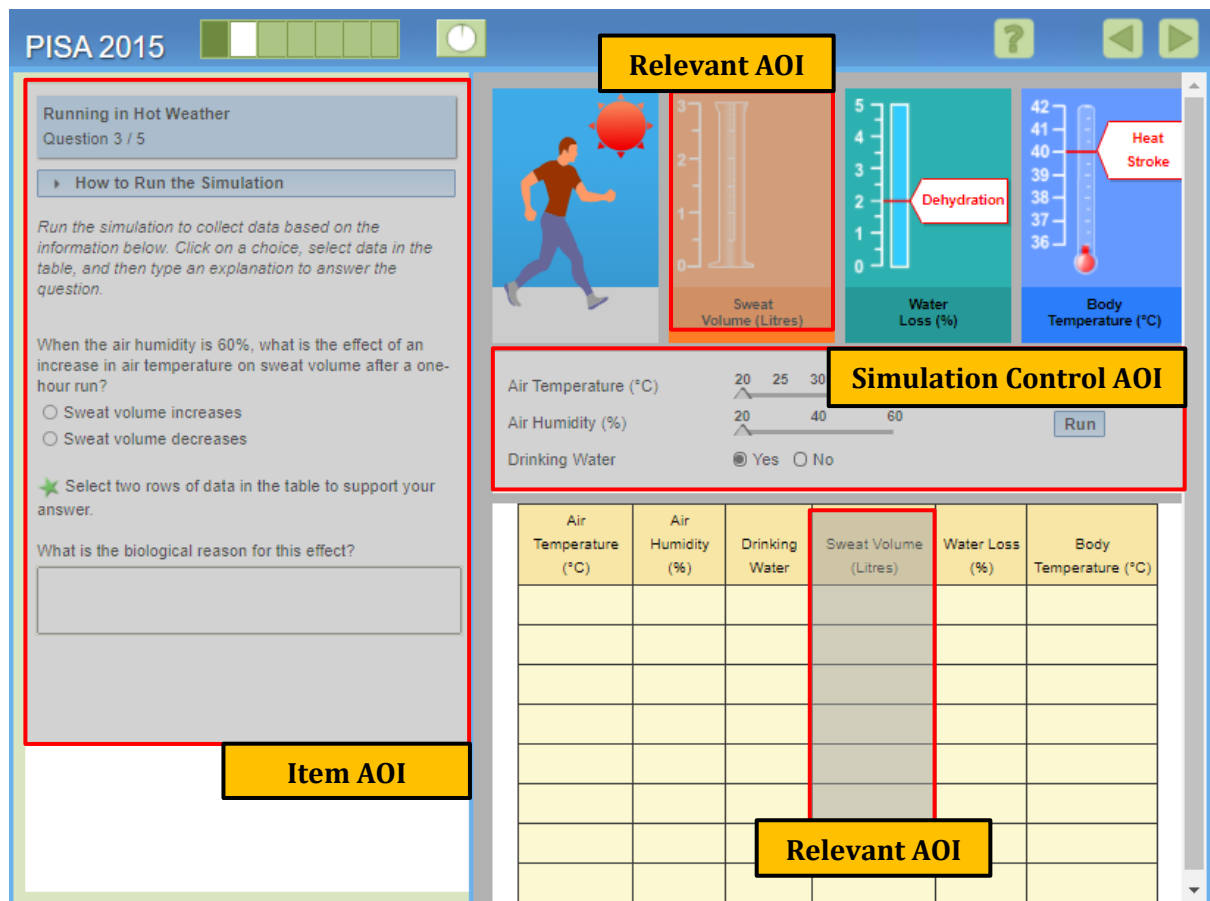


Figure 3.13 Placement of AOIs for ‘Orientation’ phase of items in Study 2

Figure 3.14 shows the size and location of the AOIs for items in this study after the first simulation was run (and for all subsequent simulations). At this point, AOIs that represented ‘Irrelevant’ information (i.e. those AOIs that contained information that did not help answer the task’s item) were identified. It should be noted that these AOIs do not ‘fill’ the content of each column. This was to ensure that there is sufficient distance between the AOIs in the table. Eye movement research has noted that there is a tendency to focus on central image regions (e.g. Gameiro et al., 2017) and recent research by Bruckmaier et al. (2019) has demonstrated that fixations in tables with figures ignore the outer edges of a cell. Therefore, the construction of these AOIs sufficiently safeguarded data quality.

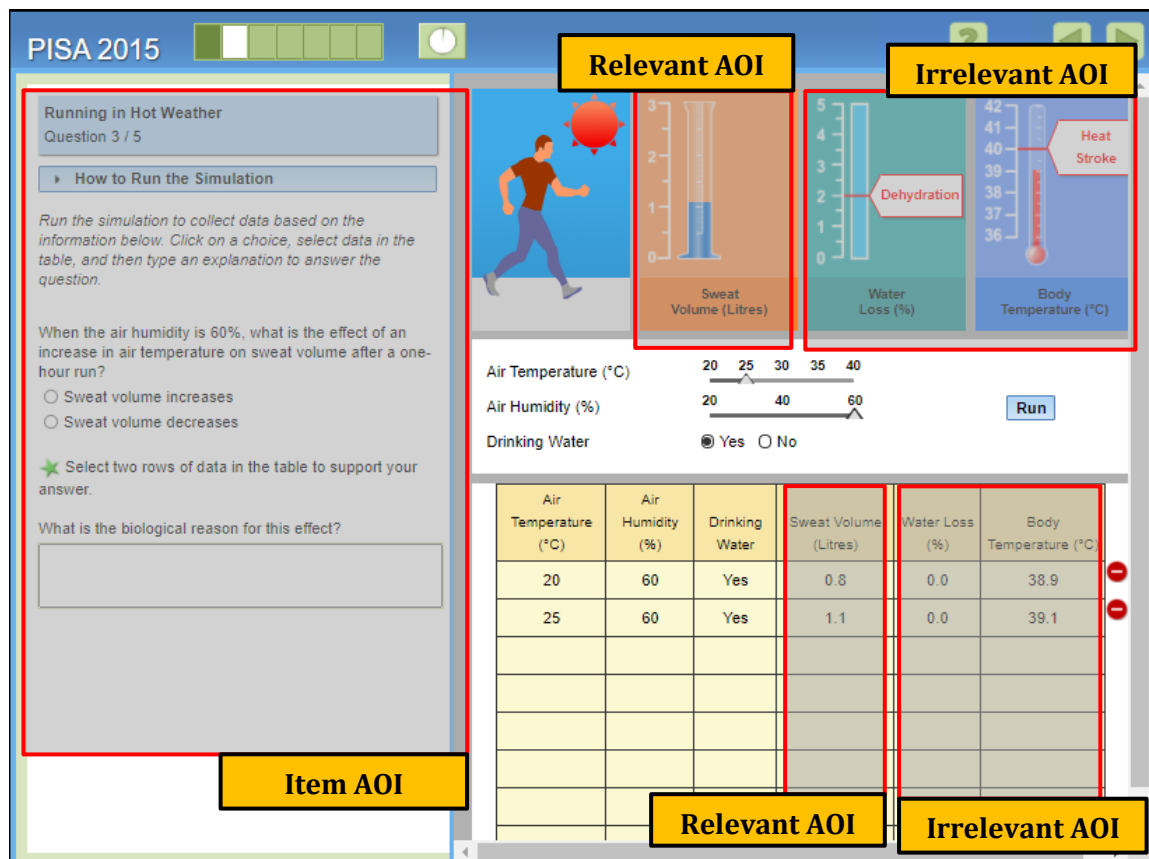


Figure 3.14 Placement of AOIs for ‘Output’ phase of items in Study 2

3.8.2.3 Eye Movement Metrics

Eye-tracking can provide a multitude of different dependent measures for analysis (Hölmqvist et al., 2011). The metrics available in the *Tobii Pro Lab* software could be related to a particular AOI (AOI-dependent) or not (AOI-independent). The former gives insight into participant behaviour on the contents of an AOI. The latter can contextualise how the participant behaved in general. In terms of eye movements, visits (portion of gaze data between the first and final fixation in an AOI), fixations (where the eye is focussed on a single point) and saccades (rapid eye movements that change the point of fixation) are the most basic units for analysis (Carter & Luke, 2020). For this study, measures of duration and location of fixations, received particular attention along with relevant visit metrics. The range of available fixation and visit metrics in the *Tobii Pro Lab* for TOIs and AOIs software are summarised in Table 3.3.

Table 3.3 Fixation and visit metrics for TOIs and AOIs

TOI metrics		
Metric Name	Description	Unit
Duration of TOI	Duration of an interval	Milliseconds
Start of TOI	Start time of an interval	Milliseconds
AOI Fixation Metrics ³⁷		
Metric Name	Description	Unit
Total duration of fixations	Total duration of fixations inside an AOI.	Milliseconds
Average duration of fixations	Average duration of fixations inside an AOI.	Milliseconds
Minimum duration of fixations	Duration of the shortest fixation inside an AOI.	Milliseconds
Maximum duration of fixations	Duration of the longest fixation inside an AOI.	Milliseconds
Number of fixations	Number of fixations occurring in an AOI.	Number
Time to first fixation	Time to the first fixation inside an AOI.	Milliseconds
Duration of first fixation	Duration of the first fixation inside an AOI.	Milliseconds
AOI Visit Metrics		
Metric Name	Description	Unit
Total duration of visit	Total duration of visits inside an AOI.	Milliseconds
Average duration of visit	Average duration of visits inside an AOI.	Milliseconds
Minimum duration of visit	Duration of the shortest visit inside an AOI.	Milliseconds
Maximum duration of visit	Duration of the longest visit inside an AOI.	Milliseconds
Number of visits	Number of visits occurring in an AOI.	Number
Time to first visit	Time to the first visit inside an AOI .	Milliseconds
Duration of first visit	Duration of the first visit inside an AOI.	Milliseconds

³⁷ Fixation metric calculations in *Tobii Pro Lab* include all fixations in a TOI or AOI. 'All' fixations include both whole fixations (they start and end within the TOI or AOI), but also partial fixations (they start before the TOI or AOI interval starts). Only whole fixations were exported for analysis. Whole fixations were defined as those which are preceded and succeeded by a saccade and were wholly contained in the TOI or AOI (Tobii AB, 2020).

Carter and Luke (2020) categorise these metrics temporally, according to the stage of processing they capture: early or late measures. Early measures used with AOIs include time-to-first-fixation (how long before an AOI is first fixated upon) and duration of first fixation. Later measures include total fixation duration (the total amount of time spent fixating a region of interest) and number of fixations in a region of interest. It is important to recognise though that late measures are not wholly independent of early measures e.g. dwell time will be influenced by time-to-first-fixation. In fact, many eye tracking metrics are highly correlated and are not independent of each other (Hölmqvist et al., 2011). This means that the selection of dependent measures for analysis should be carefully considered to reduce the risk of any possible redundancy.

In selecting measures for any eye tracking study, a clear understanding of what type of comparisons are appropriate and valid for the study were considered. For example, eye movements are influenced by a variety of factors such as visual complexity and AOI size (e.g. Nuthmann, 2017). A larger stimulus or AOI is more salient and easier to see, thus attracting more fixations than smaller stimuli. In Study 2, nearly all of the AOIs were of a similar size and identically presented to participants so this was not a concern. However, if the stimuli in an experimental eye tracking study significantly differ in their presentation, then eye movements in these conditions will also automatically differ, creating a confound if not accounted for (Carter & Luke, 2020). In Study 1B, the stimuli used in the experimental conditions *intentionally* differed on a range of visual factors e.g. the animations were more dynamic than the text and image stimuli and included an audio narration etc. Eye movements in the AOIs associated with the stimulus materials are expected to differ given the differences in visual complexity and required participant behaviour e.g. reading vs listening. Therefore, direct comparisons regarding fixation durations on the different stimuli types found in the two conditions were deemed inappropriate. In fact, Orquin and Hölmqvist (2018) outlined how total duration of fixations are a completely unsuitable metric to use when attempting to draw comparisons between experimental conditions. Instead comparisons on eye movement behaviour with the *identical* AOIs that existed across conditions were undertaken (e.g. those AOIs associated with an item's interaction space) to determine if differences in condition influenced eye movement behaviour.

For the purposes of this study, sixteen metrics in total were available and potentially relevant. However, Carter and Luke (2020, p. 55) caution against using all

possible metrics in eye tracking studies advising that ‘if multiple variables are chosen for analysis, they should not be redundant; each variable should answer a different question or provide additional information’. With this in mind and taking into consideration the work of Khedher et al. (2018), only four metrics were used to offer a summary of test-taker attentional behaviour in Study 1B and Study 2:

- Average duration of whole fixations
- Number of whole fixations
- Time to first whole fixation
- Number of visits³⁸

3.8.3 Qualitative Data

Qualitative data were collected using a c-RTA (See Appendix G). These c-RTAs were conducted with 12 participants after completing the test items for Study 1B and Study 2. Audio recordings of these interviews (233 minutes in total) were transcribed by the researcher. The textual data generated were then analysed using Braun and Clarke’s (2006) six-step framework for thematic analysis, supported by the *NVivo 12* software (QSR International, 2020).

3.8.4 Demographics

Demographic data were collected from each participant after they had completed the assessment materials. To begin, participants were asked to disclose their age to confirm that they were within the appropriate age-range for the test items. Due to the Covid-19 pandemic, the participants had not completed their State exams (Junior Certificate/Cycle) the previous year (SEC, 2020a). Alternative proxy control measures for participants’ ability were needed as a result. While the self-reporting of participants’ grades can be somewhat problematic due to issues related to social desirability and false recall, Kuncel et al. (2005) found that self-reported grades are reliable indicators of actual

³⁸ Many of these metrics were extracted to provide information on how participants engaged with different item types and stimulus presentations *without* any comparisons. Von der Malsburg and Angele (2017) demonstrated how eye tracking research is at an increased risk of Type I and Type II errors. This is because it is common to analyse multiple dependent measures in eye-tracking studies, resulting in the ‘multiple comparisons problem’ i.e. an increased probability that the null hypothesis is incorrectly rejected (Type I error). Bearing in mind the need to make appropriate *comparisons* between conditions, the stated measures were selected to address specific queries in Study 1B and Study 2.

grades. Consequently, participants were asked to record the percentages they had received in their three most recent English, mathematics and science assessments. These scores were then used to calculate an average score for the participant in that subject. These three subjects were considered most relevant to the study given the areas addressed in the research instruments. Participants were also asked to answer questions about their enjoyment of learning science and their interest in different science topics. Participants' enjoyment of learning science was determined through their responses ("strongly agree", "agree", "disagree" or "strongly disagree") to statements affirming that they generally have fun when learning science topics; that they like reading about science; that they are happy working on science topics; that they enjoy acquiring new knowledge in science; and that they are interested in learning about science. Participant interest in five broad science topics, or subjects was quantified through students' responses ("not interested", "hardly interested", "interested" or "highly interested") to topics related to the biosphere (e.g. ecosystem services, sustainability); to motion and forces (e.g. velocity, friction, magnetic and gravitational forces); to energy and its transformation (e.g. conservation, chemical reactions); to the universe and its history; and in how science can help us prevent disease. These questions were derived from the PISA 2015 (OECD, 2016a) survey materials and the current study's presentation of them in the online platform can be viewed in Appendix H.

3.9 Pilot Study

3.9.1 Pilot Study 1: Testing Platform

Prior to the main study, two small-scale pilots were conducted. The first of these related to the functioning of the testing platform required for Study 1. The purpose of the pilot study was to:

- Identify potential issues in the administration of the testing and survey materials
- Confirm the length of the testing process, and
- Collect feedback about the assessment and testing platform.

In May 2020, one school based in Leinster agreed to participate in the pilot study. The school sent the relevant link to their TY students and asked them to complete the test and to provide feedback. Eight students participated in the pilot study. The responses to

the open-ended questions asked at the end of the assessment indicated that the participants found the instructions very clear and the platform easy to navigate. The participants thought that there was significant variation in the difficulty of the test items with some noting that the items were ‘either very hard or very easy’. Interestingly, the participants suggested having an audio-visual component to the initial instruction pages. This recommendation informed the final design of the explanatory sections of the assessment. Questions on the use of test-taking accommodations (e.g. use of a reader, use of a typist) were also included. Procedures were also put in place to ensure the inclusion of any participants with reading, visual or hearing impairments³⁹.

3.9.2 Pilot Study 2: Eye-Tracking

A second pilot study was conducted to identify the most effective procedures for conducting an eye-tracking study with second-level students. Four second-level students engaged with the final testing materials in September 2020 while being tracked by the *Tobii Pro Fusion* eye-tracker. Particular attention was paid to the calibration process in order to develop a standard operating procedure for this process for the main study. Calibrations are necessary in eye-tracking research to ensure that the eye-tracker captures precise and accurate eye movement data for the individual participant by ‘mapping’ the features detected in their eye image and the physical orientation of their eye (Nyström et al., 2012). Calibration involved the participants looking at nine predefined positions in the stimulus place, as highlighted by a circle. Participants were encouraged to concentrate their gaze on their circle as best they could during the calibration process. At each point, the eye tracker captured a number of eye-image features and associated ‘their positions in the eye image with the position of the target’ (Nyström et al., 2012, p. 274). The process of calibration was fully automatic and controlled by the system. To optimise the calibration procedure a number of actions were trialled in this pilot in relation to participant positioning as per Nyström et al.’s (2012) recommendations. Those that were considered to be particularly important for the current study included the following:

- Participant was in a well-lit area in accordance with *Tobii Pro’s* (2020) guidelines.

³⁹ For example, any participant with a hearing impairment was automatically allocated to the static condition by the teacher administering the test.

- Participant was in a comfortable seating position (approximately 50-70cm from the eye-tracker, depending on their level of comfort) that they could maintain for the duration of the recording.
- The eyes of the participant were visible in the screen of the eye tracker.
- If the participant was wearing glasses, the laptop was raised to a higher level so that the eye tracker had a clear 'line of sight' to the participant's pupils, unobstructed from the glasses' frames.
- Participant glasses were cleaned prior to beginning the calibration to ensure that the infrared light had sufficient access to the pupil.

After calibration, a visual representation showing error vectors for each target location (averaged across both eyes) along with a numerical estimation of accuracy, precision and recorded valid samples was available (Figure 3.15). As seen in Figure 3.15, the white crosshairs represented the calibration target areas used by the system. Each orange circle represented a sample taken of the eye at this point on the screen. At the bottom of the screen, accuracy and precision measures were available.

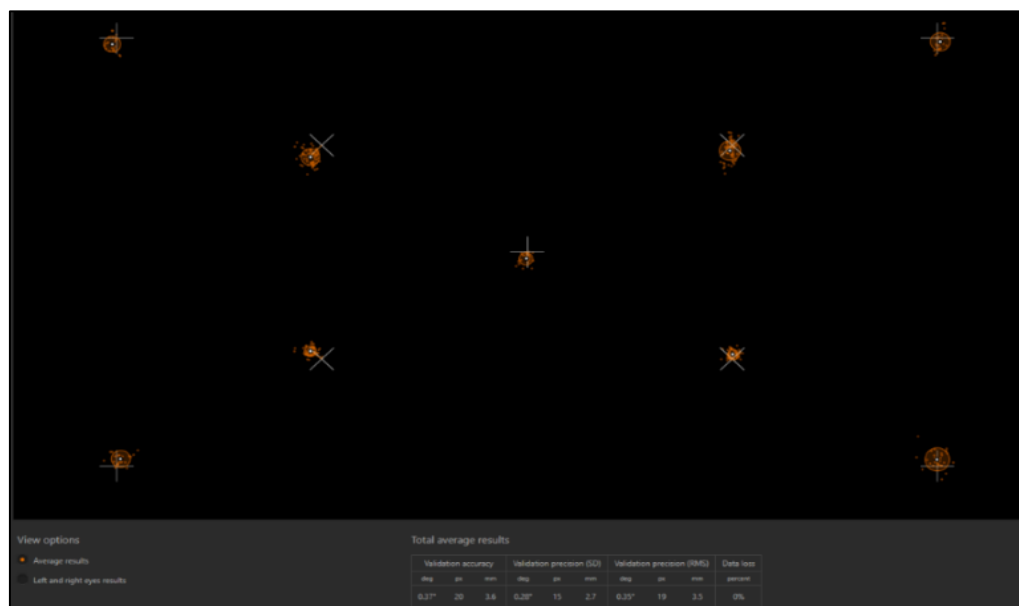


Figure 3.15 Visual representation of calibration process

A visual inspection of the calibration image was used to better understand the efficacy of the calibration process. For the calibration seen in Figure 3.15, a high level of accuracy (samples are around the target area) and precision (samples are clustered close together) is visible. While it was previously quite common to accept or reject the calibration solely on the basis of these visual representations, *Tobii Pro Lab's* measures

of calibration accuracy and precision (displayed in degrees, pixels and millimetres) were the primary source of information in deciding if recalibration was necessary. Based on the researcher's experiences in Pilot 2, a cut-off point of $.5^\circ$ was also used for both accuracy and precision as deviations of 'less than one-half of a degree are very good' in eye tracking research (Tullis & Albert, 2013) and is commonly used within the field (e.g. Krstić et al., 2018; Carter & Luke, 2020).

3.10 Main Study

3.10.1 Study 1A

Six second-level schools agreed to participate in this study between October and December 2020. As Covid-19 restrictions were ongoing, the researcher was unable to travel to the schools to administer and monitor the assessment in person. Instead, the co-operating teacher received a link from the researcher to the testing platform two days before the agreed administration day for the test. This link was then shared with students who had received permission from their guardians (see Section 3.11) to be involved in the study on administration day. Standardised administration instructions were used (Appendix I). In all schools, the test was administered in a dedicated computer room or lab using laptop computers or tablets. Once the participants completed the assent form, they were then allowed to complete the test on their screen. Participants were assigned to the two conditions by the testing platform.

One school experienced a technology failure for three of their pupils during administration. These participants were excluded from any future analyses. A problem with the testing platform's allocation procedures resulted in an over-allocation of participants to the static condition in the early stages of data collection. This was rectified as soon as possible.

3.10.2 Study 1B, Study 2 and Study 3

In October 2020, data for Study 1B, Study 2 and Study 3 were collected from participants. Participants completed the assessment on the researcher's laptop in a spare office. The following precautions were in place in light of the ongoing Covid-19 pandemic:

- All participants and the researcher wore masks.
- Hands were sanitised by participants when entering and leaving the room.

- A new pen was given to participants to fill out any forms.
- The researcher remained 1m away from the participants when they were using the laptop. A perspex glass was also between the researcher and the participant.
- The researcher had remote control of the participant's laptop (which had the eye-tracker attached).
- The laptop was fully disinfected before and after use with a specialised cleaning agent.

Before beginning Study 1B, participants' eyes were calibrated according to the procedures developed during Pilot 2. If a participants' calibration metrics exceeded the .5° cut-off point, calibration was undertaken again. If calibration could not be achieved within the agreed parameters after three calibration attempts, the participant was invited to complete the test with the eye tracker but their eye movement data were not included in any final data analyses⁴⁰. After calibration, the participants completed the assessment materials. If the participants were involved in Study 2⁴¹, a 5-minute break was provided before repeating the calibration process. The participants then completed the simulation-type items involved in Study 2. Participants who agreed to participate in Study 3 were invited to review their eye movements and comment on them after all test items in Study 2 were completed.

3.11 Ethical Considerations

This research was conducted with participants under the age of 18. As a result, they were considered a vulnerable group under DCU Ethics Guidelines (2019) and the Department of Children and Youth Affairs' guidelines (2012; 2015). However, the level of risk associated with this research project was low. The tasks associated with the research were unlikely to influence or affect the participants physically, socially, psychologically

⁴⁰ The eye movement data collected from four participants were excluded from Study 1B on the basis of these criteria. However, if the precision measure was below .30° and the accuracy measure was between .50° and .70°, a visual inspection of the participants' eye movements was undertaken to determine if the eye movements data could still be considered valid. The calibration metrics from six participants met these criteria for visual inspection. The visual inspection was conducted in line with Dalyrmple et al.'s (2018) recommendations, whereby individual data points from the calibration process were inspected. In this study, the nine calibration targets for each participant were examined to determine whether the deviation from these targets was acceptable. Data collected from two of these participants were excluded from the data set after this inspection. The remainder was included on the basis of this visual inspection. The results of the researcher's visual inspection were supported by the *Tobii Pro* research support team.

⁴¹ Due to the timing of class periods, not all participants involved in Study 1B were available for Study 2.

or spiritually. It did not deal with any sensitive issues and topics like drug abuse and the actions required by participants were consistent with their day-to-day school activity. The disclosure of any personal data (e.g. age, test scores) was optional and the eye movement data were collected using non-invasive equipment. The researcher was a fully garda-vetted member of the Irish Teaching Council and was vetted again prior to commencing this research project. The researcher also completed training on Children First Guidelines in line with DCU Ethics Guidelines (2019). Ethical approval was obtained from the DCU Ethics Committee in December 2019 (DCUREC/2019/208) and again in May 2020 (to reflect changes to the project cause by the Covid-19 pandemic; Appendix J).

Two weeks in advance of the administration date, an online guardian consent form (with an audio-visual plain language statement) was sent by the researcher to the co-operating teacher in the school (hosted on the e-Survey Creator platform). The co-operating teacher then distributed this to the parents/guardians of the Transition Year students. Using a secure link, the co-operating teacher accessed the list of those with guardian consent the day before administration day. Only those students with guardian consent were involved in the study. Participants also completed an assent form.

3.12 Data Analysis

A consistent approach to data analysis was applied throughout the research study. Quantitative data (Study 1A, Study 1B, Study 2) were analysed using *SPSS 28* (Statistical Package for the Social Sciences; 2020). Analysis began with a preliminary investigation of the relevant variables within the data set. Descriptive statistics, summarised in the form of valid percents, provided information on the distribution, central tendency and dispersion of each variable. When testing any hypotheses using inferential statistics, parametric tests were applied. However, non-parametric approaches to statistical testing were used with any variables that appeared to violate the assumptions of linearity, normality or homoscedasticity⁴². The *p*-values for parametric and non-parametric tests were set at $p=0.05$ but Bonferroni corrections were applied to control for Type I errors when multiple significance tests were carried out on the same dependent variable. To identify the magnitude of an observed effect, effect sizes were also calculated where

⁴² Given the relatively small size of the samples used in each study, the normality of a variable was assessed by a visual inspection of the relevant histograms, scatterplots and box plots and by obtaining the relevant skewness and kurtosis *z*-scores for each variable. Skewness and kurtosis *z*-scores with an absolute value greater than 1.96 were used to identify variables that were not normally distributed (Field, 2018).

appropriate. A range of effect sizes were calculated. These have been summarised in Table 3.4 alongside their descriptors (Cohen, 1988; Field, 2018).

Table 3.4 Descriptors for reporting and interpreting effect sizes

Descriptor	Cohen's d	r^*	φ	η^2
Small	0.20	.10	0.10	0.01
Moderate	0.50	.30	0.30	0.06
Large	0.80	.50	0.50	0.14

* These descriptors were used for all correlations e.g. Spearman's ρ , Kendall's τ

The qualitative data collected for Study 3 were analysed using Braun and Clarke's (2006) six-step framework for thematic analysis. Table 3.5 outlines how each step of Braun and Clark's (2006) framework was applied. The *NVivo 12* software (QSR International, 2020) programme supported this process.

Table 3.5 Braun and Clarke's (2006) six-step framework for thematic analysis

Step	Description/ Actions
Familiarisation	Audio recordings (233 minutes) were transcribed. These transcripts were read multiple times to facilitate familiarisation.
Initial Coding	The data were then organised into smaller 'chunks' or 'codes' in a meaningful and systematic way (Appendix K).
Theme Search	The initial codes were then organised into broader themes. These were descriptive in nature (Appendix L).
Theme Review	The initial themes were modified and refined and the data contained within each theme were reviewed (Appendix M).
Theme Definition	Efforts were made to '...identify the essence of what each theme is about' (Braun & Clarke, 2006, p. 92). Relationships between and within themes were also identified.
Write-up	The final thematic framework was finalised.

3.13 Summary

This chapter presented the methodology of this study. It provided information about the conceptual framework and how it informed the experimental design, through which simulation-type items and items with static and dynamic stimuli were investigated and compared. The measures used, the sampling and the procedures followed in the study were discussed in detail along with the efforts undertaken to create a valid eye-tracking study. In the next chapter, the results of this research study are presented.

Chapter 4

Results

4.1 Introduction

This chapter presents the findings of the research in four parts. The first part details the outcomes of Study 1 according to the research questions outlined in the previous chapter. In this study, post-primary students completed a TBA of scientific literacy that used either dynamic or static multimedia stimuli. For Study 1A, participants' performance on these TBAs were compared. Eye movement data were also collected from a small subset of this sample for Study 1B to better understand test-takers' attentional behaviours while completing items with differing formats and multimedia stimuli. Study 2 also involved the analysis of eye movement data from those who participated in Study 1B but these data related to their attention and performance on five additional simulation-type items. The final section of this chapter contains the analyses of the qualitative data derived from the cognitive interviews conducted for Study 3. Figure 4.1 summarises the data collected for each study.

	Study 1A	Study 1B	Study 2	Study 3
Schools (n)*	6	1		
Participants (n)**	251	33	24	12
Quantitative Data				
<i>Product Data</i>				
Test Scores	Y	Y	Y	
<i>Process Data</i>				
Eye Movement		Y	Y	
Time-on-Task			Y	
Number of Simulations			Y	
Qualitative Data				
Interviews				Y

* The school involved in Study 1B, Study 2 and Study 3 is the same. It is a sub-sample of the schools from Study 1A.

** The participants of each study are a sub-sample of the participants involved in the preceding study.

Figure 4.1 Summary of Data Collected (by Study)

4.2 Study 1A

Participants in Study 1A were randomly assigned to take either the dynamic (animated) or static (text-image) version of a TBA for scientific literacy. These two groups formed the experimental and control groups of this research study⁴³. The sample's demographic details, including evidence regarding the equivalence of the experimental and control groups, will first be provided. The key findings of this study will then be presented according to the research questions outlined in Chapter 3.

4.2.1 Demographics

A total of 251 Irish second-level students participated in Study 1A. Disclosure of their demographic data were optional for the participants, with 96% ($n=237$) providing these data. Demographic data included the participants' ages and their performance on their three most recent school assessments in English, mathematics and science (Table 4.2). The average age of the participants in Study 1A was 16 years. There were no statistically significant differences in participants' self-reported average performance in school-based English [$t(235)=0.26, p=.80$], mathematics [$t(235)=-1.15, p=.53$] or science [$t(235)=-0.17, p=.46$] assessments. Participants who completed the dynamic version of the test did not appear to be significantly different to those who completed the static version in terms of the relevant background characteristics (Table 4.1).

Table 4.1 Demographic Details: Study 1A

	Static (Text-Image) $n=130$		Dynamic (Animations) $n=107$	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Age	15.6	0.5	15.6	0.6
Average Performance in English Assessments (%)	68.2	12.4	67.8	12.3
Average Performance in Mathematics Assessments (%)	63.0	16.5	65.6	18.2
Average Performance in Science Assessments (%)	71.0	15.2	71.3	15.4

⁴³ The static condition was considered the 'control' group as the items in this condition were replicated from the original PISA 2015 (OECD, 2018; 2017) items.

Participants were also asked to disclose their thoughts on science as a subject using items from PISA 2015 (OECD, 2016a; 2016b). Patterns of responses to these questions were highly similar across the experimental and control groups. For the most part, the sample in Study 1A did not seem to be positively predisposed to the subject (Table 4.2). The majority of students disagreed with almost all of the statements relating to interest in and enjoyment of the subject, with the exception of ‘I like reading science-related articles and books’, which just over 60% of students endorsed.

Table 4.2 Percentages of students agreeing/disagreeing with statements about their enjoyment of learning science

	Agree/ Strongly Agree	Disagree/ Strongly Disagree
	%	%
I like reading science-related articles and books.	61.2	38.8
I am interested in learning about science	24.5	75.5
I generally have fun when I am learning about different topics in science class.	22.4	77.6
I am happy working on science-based activities.	17.3	82.7
I enjoy acquiring new knowledge in science.	16.0	84.0

n=237

Participants were also asked about the extent to which they are interested in five science topics: ‘The Biosphere’ (e.g., ecosystems and sustainability), ‘Motion and Forces’ (e.g., velocity, friction, magnetic and gravitational forces), ‘Energy and its Transformation’ (e.g. conservation, chemical reactions), ‘The Universe and its History’, and ‘how science can help us prevent disease’. Participants in Study 1A were mostly interested in ‘how science can help us to prevent disease’ (92.0%). It should be noted that at the time of data collection the COVID-19 pandemic was ongoing and may have influenced participants’ responses. ‘Motion and Forces’ was the least popular science topic among this sample with 50.2% indicating that they had ‘no interest’ in it (Figure 4.2).

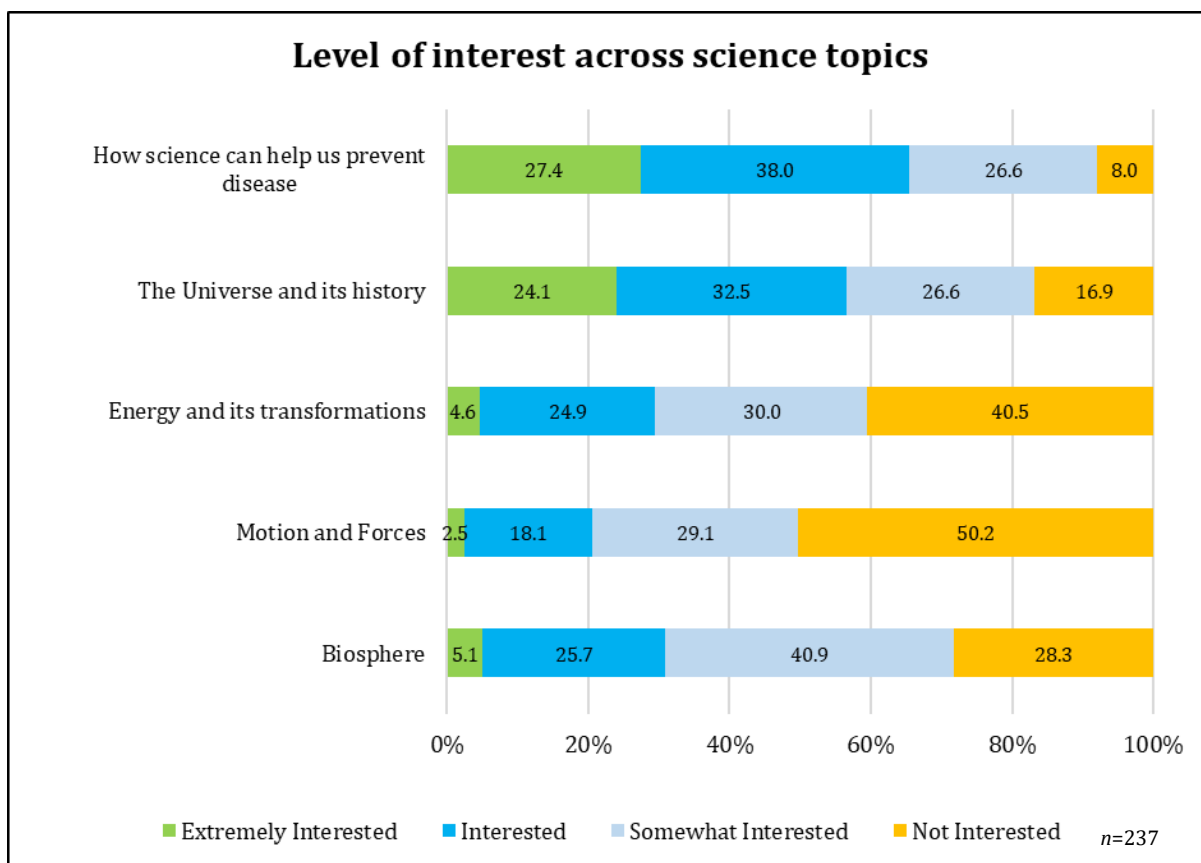


Figure 4.2 Percentages of students who agreed/disagreed with various statements about their interest in different science topics

4.2.2 Research Question 1 (RQ1): Do different multimedia stimuli (e.g. images, animations) affect test-taker performance in a TBA of scientific literacy?

The first set of research questions (RQ1) attempted to determine the impact of different multimedia stimuli on test-taker performance. Analyses were conducted to better understand if the multimedia stimulus used in each condition influenced the overall performance of test-takers. Based on previous literature (e.g. Wu et al., 2015), an investigation into the possible effect of multimedia stimulus on test-takers' performance where different levels of prior knowledge were present was also conducted. The performance of each item across the two experimental conditions was also considered.

4.2.2.1 RQ1a: Is the performance of test-takers on items in a TBA of scientific literacy affected by the type of multimedia stimulus used?

The findings contained in Table 4.3 show that, overall, participants performed relatively well on the TBA of scientific literacy, getting more than half of all the items correct ($M=53.1$, $SD=19.3$). An independent-samples t -test found no statistically significant difference in scores between those who took the animated ($M = 54.0$, $SD = 20.1$) and the text-based TBA ($M = 52.5$, $SD = 18.7$), $t(249) = -0.55$, $p = .58$. The effect size calculated for this was small; $d=-0.1$. A statistically significant difference in the amount of time it took the participants to complete the TBA, was noted along with a relatively large effect size. Participants spent more time in completing the dynamic version of the test ($M = 16.0$, $SD = 5.7$) than the static one ($M = 13.4$, $SD = 4.8$); $t(246) = -4.14$, $p = <.001$, $d=-0.53^{44}$ (Table 4.3).

Table 4.3 Performance of participants involved in Study 1A

	Test Performance (%)		Test Duration (mins)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Overall	53.1	19.3	14.5	5.4
<i>Conditions</i>				
Static (Text, Image; $n=141$)	52.5	18.7	13.3	4.8
Dynamic (Animated; $n=110$)	53.9	20.	16.0	5.7

$n=251$

Frequencies of correct answers on individual items by participants are shown in Table 4.4. In line with the findings of the previous t -test, the relative frequency with which participants got individual items correct did not appear to vary at a statistically significant level across conditions, with one exception in Item P4 which was more likely to be answered correctly when presented dynamically; $\chi^2(1) = 4.07$, $p=.04$. However, the effect size calculated for this difference was not practically meaningful; $\phi=0.14$.

⁴⁴ The sample size for this comparison was $n=248$ as three participants did not 'log off' from the testing platform (times were in excess of 50 minutes). Some participants (5 in total) in this group spent less than 5 minutes completing the TBA. However, as they did complete the CR items on the test with a sufficient level of accuracy, they were considered to still have relevant data to contribute to the study and were included in the analysis.

Table 4.4 Frequencies of correct answers on individual items: Static, Dynamic

Item	Static (<i>n</i> =141)	Dynamic (<i>n</i> =110)	$\chi^2(1)$	<i>p</i>	ϕ
M1	48%	40%	1.38	.24	-0.08
M2	66%	62%	0.30	.59	-0.04
M3	90%	86%	0.85	.36	-0.07
M4	65%	66%	0.03	.96	0.01
F1	4%	3%	0.92	.76	-0.04
F2	70%	65%	0.67	.41	-0.06
F3	65%	67%	0.04	.84	0.02
P1	13%	18%	1.02	.31	0.08
P2	53%	58%	0.60	.44	0.06
P3	57%	49%	1.16	.28	-0.08
P4	48%	62%	4.07	.04*	0.14
G1	28%	36%	1.47	.23	0.09
G2	61%	72%	2.75	.10	0.11
G3	77%	83%	1.07	.30	0.08
G4	43%	42%	0.00	1.00	-0.01

n=251* *p*<.05

4.2.2.2 RQ1b: Is the performance of test-takers in a TBA affected by the type of multimedia stimulus used when their previous levels of prior knowledge are considered?

A two-way independent ANOVA was conducted to explore the impact of condition and prior level of science knowledge on performance in the study's TBA. Test-takers' prior science knowledge was measured by calculating the average score of their three most recent class-based science tests (self-reported). Participants in both conditions were subsequently categorised into three subgroups according to their level of prior knowledge: high, moderate and low. Participants who had achieved an average score greater than or equal to 75% in their class-based science tests were considered to have 'high' levels of prior knowledge. Those who scored between 55% and 75% were classified as having 'moderate' levels of prior knowledge. Those who scored less than or equal to

54% had 'low' levels of prior knowledge⁴⁵. Table 4.5 summarises the number of participants within each category of prior knowledge across the conditions.

Table 4.5 Scores of participants in Study 1A according to condition and levels of prior science knowledge

	Low			Moderate			High		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Static (Text, Image)	20	34.3	12.7	49	47.2	16.5	61	62.7	15.3
Dynamic (Animated)	12	29.4	15.2	45	48.3	14.2	50	66.3	17.1

n=235

Unsurprisingly, there was a significant main effect for level of prior knowledge with a large effect size; $F(2, 231)=62.86$, $p<.001$, $\eta^2=.35$. Post hoc tests (with a Bonferroni correction) revealed that test scores on the TBA were significantly higher for those with high levels of prior science knowledge when compared with those with moderate ($p<.001$) or low levels ($p<.001$). Those with moderate levels of prior science knowledge scored significantly higher than those with low levels of prior knowledge ($p<0.001$). The interaction effect between condition and level of prior science knowledge was not statistically significant, $F(2, 231)=0.88$, $p=0.42$ but a small effect size was noted, $\eta^2 = .01$ ⁴⁶.

4.2.2.3 RQ1c: Does the type of multimedia stimulus used affect key item statistics?

To determine the consistency of the items used in the TBA, reliability analyses were conducted for the entire data set and for each condition. These are summarised in Table 4.6. The reliability values of these items as a measure for scientific literacy were approaching acceptable levels according to Cohen et al.'s (2011) recommendations

⁴⁵ Levels of prior knowledge were classified using a modified version of the Junior Cycle standards (SEC, 2020b). 'High' levels of prior knowledge in this study aligned with the 'Higher Merit' ($\geq 75\%$) or 'Distinction' ($\geq 90\%$) grade descriptors for Science. 'Moderate' levels of prior knowledge were identified based on the 'Merit' ($\geq 55\%$) or 'Achieved' ($\geq 40\%$) descriptors. Participants were classified as having 'low' levels of prior knowledge if their average scores fell within the percentage bands for the 'Achieved' ($\geq 40\%$), 'Partially Achieved' ($\geq 20\%$) or 'Not Graded' ($\geq 0\%$) grades.

⁴⁶ The eta squared (η^2) effect size is recommended for use with two-way ANOVAs (Pallant, 2007). Cohen's (1988) criterion respectively classes .01, .06 and .14 as small, medium and large effect sizes.

where values greater than 0.70 are recommended. For the overall data set ($n=251$), a reliability of 0.69 was calculated based on the 15 items. For each condition, relatively similar Cronbach's α were calculated with a slightly higher level emerging for those that viewed animated stimuli (Cronbach's $\alpha = 0.71$) compared to those in the other condition (static; Cronbach's $\alpha = 0.66$).

Table 4.6 Values for Cronbach's alpha in Study 1A

	<i>n</i>	α
Overall	251	0.69
<i>Conditions</i>		
Static (Text, Image)	141	0.66
Dynamic (Animated)	110	0.71

Item difficulties, and their subsequent rank order, were also calculated for each condition (Table 4.7). A paired-samples t-test where the mean difficulty of the 'control' (static) items was compared to the mean difficulty of the dynamic items (which were modified from the original static items) was subsequently conducted. No statistically significant or practical differences between the means were noted; $t(14) = -0.80$, $p = .44$ (two-tailed), $d=0.10$. Furthermore, the rank ordering of the item difficulties was almost identical across the two conditions, with some exceptions (e.g. P3, P4, G2). No obvious pattern could explain these variations in item difficulties between conditions. For example, Item P4 was an easier item to get correct for participants in the dynamic condition. The inverse was true for Item P3. The grey shading in Figure 4.7 indicates those items that had noticeable differences in their difficulty indices between the two conditions.

Table 4.7 Rank order of items by difficulty across conditions

Item	Item Difficulty			Rank Order		
	Overall	Static	Dynamic	Overall	Static	Dynamic
M1	0.45	0.48	0.40	11	11	12
M2	0.64	0.66	0.62	7	4	7
M3	0.88	0.90	0.85	1	1	1
M4	0.66	0.65	0.66	5	5	5
F1*	0.04	0.04	0.03	15	15	15
F2	0.68	0.70	0.65	3	3	6
F3	0.66	0.65	0.67	4	6	4
P1	0.15	0.13	0.18	14	14	14
P2	0.55	0.52	0.58	8	9	9
P3	0.53	0.57	0.49	10	8	10
P4	0.54	0.48	0.62	9	10	8
G1	0.32	0.28	0.36	13	13	13
G2	0.66	0.61	0.72	6	7	3
G3	0.79	0.77	0.83	2	2	2
G4	0.42	0.43	0.42	12	12	11
Average	0.53	0.53	0.49			

* This item was one of the most difficult ones included in PISA 2015 (OECD, 2017). Of the half million students involved in PISA 2015, only 5.5% (*SD*: 0.1) got this item correct. The item difficulty calculated for this item in PISA 2015 was 1.31 (OECD, 2017).

Item discriminations were also calculated. The overall average point biserial discrimination index for an item was 0.29, with a range from 0.01 to 0.44. For static items, the average biserial discrimination index was 0.27 (Range: 0.01 – 0.44) and for dynamic items it was 0.31 (Range: 0.00 – 0.44). There was no statistically significant difference between the discrimination indices calculated for both conditions; $t(14) = -1.69$, $p = .11$ (two-tailed). The effect size was also negligible; $d = -0.06$. Inspection of the discrimination indices on an item-by-item basis however, revealed that there were some discrepancies between conditions e.g. P1, P4, G1, G2, G3. These are highlighted in Table 4.8. The grey

shading in Table 4.8 indicates those items that had noticeable differences in their discrimination indices between the two conditions.

Table 4.8 Item discriminations across conditions

Item	Overall	Static	Dynamic
M1	0.20	0.19	0.22
M2	0.40	0.42	0.39
M3	0.24	0.23	0.25
M4	0.36	0.38	0.33
F1	0.01	0.01	-0.00
F2	0.36	0.34	0.39
F3	0.40	0.43	0.37
P1	0.30	0.21	0.40
P2	0.14	0.11	0.17
P3	0.36	0.33	0.40
P4	0.34	0.27	0.42
G1	0.37	0.43	0.30
G2	0.13	0.07	0.20
G3	0.27	0.18	0.38
G4	0.44	0.44	0.44
Average	0.29	0.27	0.31

As seen in Table 4.8, Item P1 had a discrimination index of 0.21 in the static condition and 0.40 in the dynamic condition. This item was better at distinguishing between those with high and low levels of prior science literacy knowledge when it was represented dynamically. This was replicated in Items P4, G2 and G3. The opposite was true for Item G1 where the item in its static form (0.43) was more successful in measuring test-taker knowledge than its dynamic counterpart (0.30). Figure 4.3 illustrates how item discriminations differed between the control and experimental conditions.

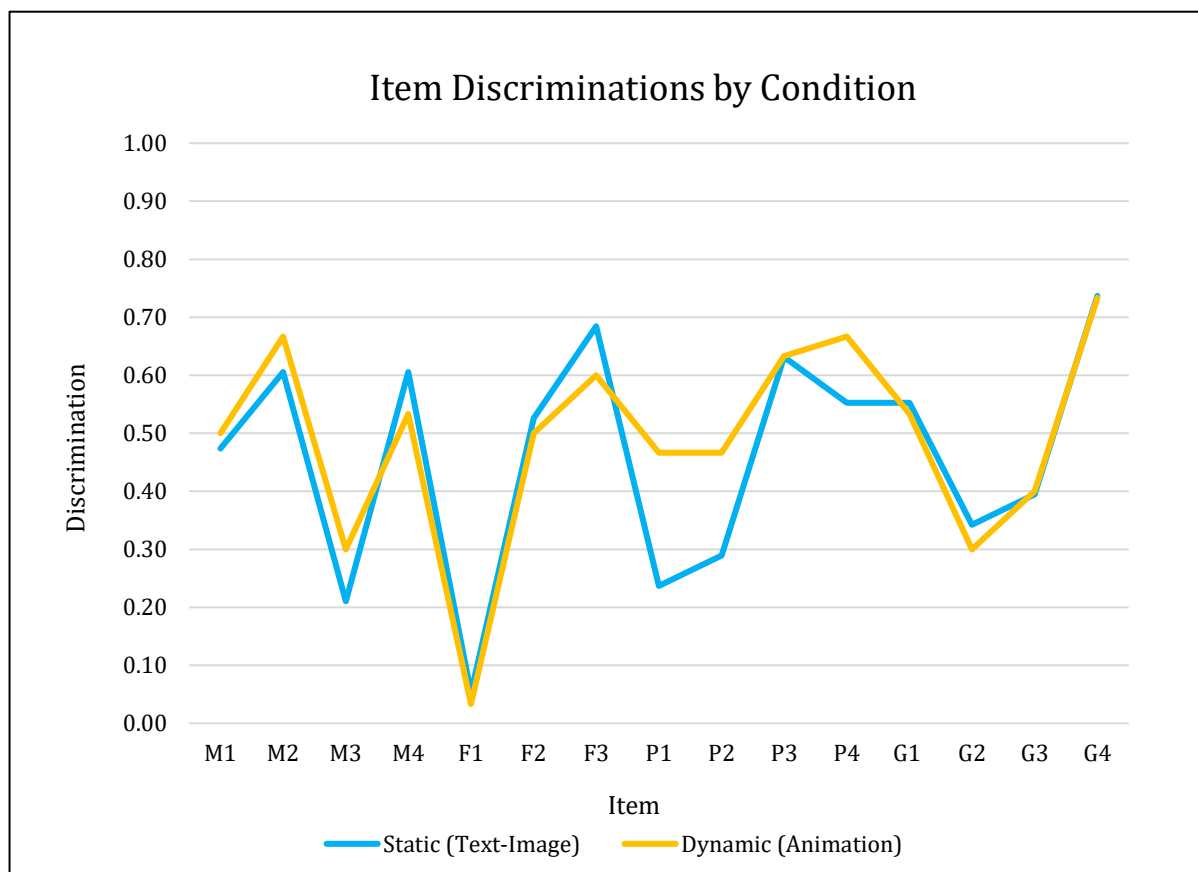


Figure 4.3 Comparison of item discriminations across conditions

4.2.3 Summary: Study 1A

Study 1A investigated if different multimedia stimuli (e.g. images, animations) can affect test-taker performance. Using an experimental approach, this study compared test-taker performance on a static and dynamic version of a TBA of scientific literacy. No differences between test-taker performance according to these conditions were found (RQ1a). When previous levels of prior knowledge in science were taken into account, the performance of test-takers did not differ to an extent that was statistically significant (RQ1b). Furthermore, item performance did not appear to be systematically influenced by condition. However, some idiosyncrasies in relation to this were noted for certain items (RQ1c).

4.3 Study 1B

While completing Study 1A, eye movement data were also collected from a sub-sample of participants to provide greater insight into test-takers' attentional behaviours while completing each item in the TBA. Taking into consideration the results from Study 1A, where there were no statistical or practical differences noted between test-takers across conditions, a large scale comparison of attentional behaviours on an item-by-item basis was considered unnecessary. Instead, only those items with the largest differences in at least one of the item performance indicators were investigated further; specifically Items P1-P4 and Items G1-G3 (see Appendix C)⁴⁷. The key findings of this study are now presented according to the research questions outlined in Chapter 3.

4.3.1 Demographics

A total of 33 second-level students formed the sub-sample that participated in Study 1B⁴⁸. The participants were randomly assigned by the testing platform to take either the dynamic ($n=16$) or static ($n=17$) version of the TBA for scientific literacy. Table 4.9 summarises the participants' ages and their performance (in percentages) on their three most recent school assessments in English, mathematics and science.

Table 4.9 Demographic Details: Study 1B

	Static (Text, Image) $n=17$		Dynamic (Animations) $n=16$	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Age	15.5	0.3	15.7	0.2
Average Performance in English Assessments (%)	64.7	11.8	69.7	10.0
Average Performance in Mathematics Assessments (%)	61.3	14.3	66.3	16.6
Average Performance in Science Assessments (%)	69.3	12.9	75.7	12.6

⁴⁷ While Item P2 did not meet this criterion, it was still considered relevant for review given its relationship to the other items selected for further investigation i.e. part of the same unit. Furthermore, the discrimination indices computed indicated that the item (which was the second most difficult in the test) was better at discriminating between test-takers in the animated condition. In contrast, G4 was not considered for inclusion as the differences between conditions on key item statistics was negligible; no differences in item difficulties across conditions and a difference of 0.01 for item discriminations.

⁴⁸ Forty participants were recruited for the study. However, eye movement data from seven participants were excluded from the analysis as their eyes were unable to calibrate with the eye-tracker machine.

As in Study 1A, the average age of the participants in Study 1B was 16. There were no statistically significant differences in participants' self-reported average performance in school-based English [$t(31)=-1.29, p=.21$], mathematics [$t(31)=-0.93, p=.36$] or science [$t(31)=-1.44, p=.16$] assessments. In terms of their enjoyment of science as a subject and the areas they found most interesting, the views of this sub-sample, by and large, reflected the findings from Study 1A. There was also no statistically significant difference in scores between those who took the animated ($Md = 10.0$) and the text-based version ($Md = 9.0$), $U=142.00, z=0.22, p=.85, r=-.04$. In contrast to Study 1A, there was no statistically significant difference between conditions on the amount of time test-takers took to complete the TBA; $U=146.00, z=0.36, p=.74, r=-.06$.

4.3.2 Research Question 2 (RQ2): How do different multimedia stimuli (e.g. images, animations) affect the attentional behaviour of test-takers in TBAs?

The second research question (RQ2) aimed to identify the impact, if any, of different multimedia stimuli on test-taker attentional behaviour using three eye movement metrics: visit count⁴⁹, average duration of whole fixations⁵⁰ and proportion of fixations⁵¹. These metrics are considered 'late' eye movement measures (Carter & Luke, 2020). Taking into consideration the impact that this may have on Type I errors, each metric will address a different aspect of attentional behaviour (Hölmqvist et al., 2011). Furthermore, to ensure that only appropriate comparisons were undertaken, eye movement data were only examined in relation to an item's interaction space (see Section 3.8.2.2 for further details).

4.3.2.1 RQ2a: Does the number of visits to an item's interaction space differ according to the multimedia object used?

A visit begins when a test-taker first fixates on an AOI and ends when the test-taker fixates on something outside that AOI (Hölmqvist et al., 2011). When the test-taker returns to the AOI it is counted as another visit. Visit counts represent how often a test-taker returned to look at a particular AOI and are independent of the number of fixations

⁴⁹ Visit counts summarise how often someone returned to an AOI (Area of Interest) (Alemdag & Cagiltay, 2018).

⁵⁰ Average duration of fixations refers to how long a fixation lasted, on average, in an AOI (Alemdag & Cagiltay, 2018). This was calculated by eye-tracking software (Tobii Pro, 2020).

⁵¹ This refers to what proportion of the total fixations that occurred during an item's presentation was allocated to a particular AOI (Luke & Asplund, 2018).

associated with an AOI. The number of visits to the interaction space of an item between both conditions were compared using a series of Mann-Whitney U tests. As illustrated by Table 4.10, test-takers in the dynamic condition were more likely to leave and then return to the interaction space of an item than those in the static condition. With the exception of Item P4, this appears to have been a relatively consistent behaviour of those on the dynamic condition. Furthermore, this difference in behaviour was statistically significant for four items (Items P1, P2, G2 and G3). Moderate to large effect sizes were noted for each of these differences (Table 4.10). It should be noted that in line with Field's (2018) recommendations for non-parametric tests, medians are being reported.

Table 4.10 Number of Visits to Interaction Space by Condition

Item	<i>U</i>	<i>z</i>	<i>p</i>	<i>r</i>	Static	Dynamic
					<i>n</i> =17	<i>n</i> =16
					Md	Md
P1	203.00	2.43	.02*	.43	7.0	10.5
P2	191.00	2.00	.05*	.35	4.0	7.0
P3	129.50	-0.24	.82	.04	4.0	4.5
P4	137.00	0.36	.99	.06	3.0	3.0
G1	184.50	1.76	.08	.31	9.0	12.0
G2	211.00	2.72	.01*	.47	8.0	12.0
G3	209.50	2.66	.01*	.46	5.0	8.0

n=33

* *p*<0.05

4.3.2.2 RQ2b: Does the average duration of whole fixations in the interaction space of an item differ according to the multimedia stimulus used?

To determine if test-taker attentional behaviour differed while looking at the interaction space of an item, average fixation durations were examined. Average fixation durations are a measure of how long, in milliseconds, fixations lasted in a particular AOI

(Carter & Luke, 2020)⁵². Comparisons of the average duration of fixations between conditions are summarised in Table 4.11.

Table 4.11 Average Duration of Whole Fixations on Interaction Space by Condition

Item	<i>U</i>	<i>z</i>	<i>p</i>	<i>r</i>	Static	Dynamic
					Md (ms)	Md (ms)
P1	177.00	1.48	.15	.26	219.0	253.5
P2	222.00	3.10	.00*	.54	216.0	279.0
P3	199.00	2.27	.02*	.40	269.0	309.5
P4	174.50	1.39	.17	.25	221.0	274.5
G1	174.00	1.37	.18	.29	204.0	232.0
G2	164.50	1.03	.31	.18	218.0	248.0
G3	177.00	1.48	.15	.26	213.0	256.5

* $p < 0.05$

As seen in Table 4.11, the average duration of a fixation in the interaction space of an item was longer for test-takers in the dynamic condition. Two items, had a statistically significant difference in the average duration of fixations between those who took the static and dynamic TBAs – Items P2 and P3. Both were complex SR items where participants had to select the correct words to complete a sentence. The median duration of fixations in the interaction space of Item P2 by participants in the dynamic conditions was 279.0ms. For those in the static condition it was 216.0ms. The difference in durations between these two conditions had a large practical difference based on the effect size calculated ($r=.54$). Similarly, test-takers who completed the animated version of Item P3 had, on average, longer fixation durations in the interaction space of an item than those in the static condition. This had a moderate effect size ($r=.40$) associated with it. Although comparisons of average fixation durations across conditions for the other items did not indicate any statistically significant differences, small to moderate effect sizes were noted for each.

⁵² Average fixation durations for each participant were calculated by the Tobii Pro software (Tobii Pro, 2020). The sequence of raw gaze points that make up a single fixation, where the estimated velocity is below the velocity threshold set in the eye-tracker's gaze filter, were aggregated by the software to calculate the average fixation durations.

4.3.2.3 RQ2c: Does the proportion of fixations to the interaction space of an item differ according to the multimedia object used?

Proportions of fixations to the interaction space of an item were calculated by relating the number of whole fixations on this area to the total number of whole fixations for that item (as per Luke & Asplund, 2018). As shown by the data in Table 4.12, test-takers in the dynamic condition (with one exception), allocated more of their attention to the interaction space of an item than those in the static condition. For Items P2 and G3, this difference in behaviour was statistically significant, with large effect sizes noted (ranging from 0.42 - .58). However, in the case of Item P2, participants in the static condition allocated more of their fixations to the interaction space than those in the dynamic condition.

Table 4.12 Proportion of Whole Fixations on Interaction Space by Condition

Item	<i>U</i>	<i>z</i>	<i>p</i>	<i>r</i>	Static	Dynamic
					Md (ms)	Md (ms)
P1	92.00	-1.59	.12	.28	0.35	0.48
P2	43.00	-3.35	.00*	.58	0.89	0.69
P3	133.00	-0.12	.93	.16	0.74	0.80
P4	158.00	0.79	.44	.14	0.59	0.89
G1	159.50	0.85	.40	.15	0.06	0.07
G2	172.00	1.30	.20	.23	0.23	0.31
G3	202.00	2.38	.02*	.41	0.40	0.53

**p*<0.05

4.3.3 Summary: Study 1B

Study 1B used eye movement data to investigate if different multimedia stimuli can affect test-taker behaviour. Using their performance on key item statistics as a criterion, seven items from Study 1A were examined in Study 1B. Test-taker attentional behaviour between conditions on these items were compared using three eye-tracking metrics: Number of visits, average duration of fixations and proportion of fixations. While statistically significant differences between test-takers on these variables were not always found, the effect sizes calculated do indicate that different multimedia stimuli can

affect how test-takers interact with the interaction space of an item in a TBA. For example, test-takers in the dynamic condition were more likely to leave and then return to the interaction space of an item than those in the static condition (RQ2a). While attending to the interaction space, the average fixation duration of test-takers in the dynamic condition were consistently longer (RQ2b). With one exception (Item P2), the distribution of test-takers' fixations on the interaction space of an item was greater in the dynamic condition (RQ2c).

4.4 Study 2

For Study 2, participants engaged with a series of inquiry assessment tasks while being monitored by the eye-tracker. For each task, participants had to co-ordinate the effects of multiple variables to run a range of simulations that would help them to answer the item for that task. Study 2 explored how basic behavioural measures and eye movement data can provide insight into students' inquiry performances. Test-taker performance on these tasks are now presented alongside relevant item statistics. Behavioural measures (e.g. time-on-task, number of simulations run) and their relationship to task performance were then investigated. Early and late measures of eye movement (e.g. time-to-first-fixation, number of fixations) will subsequently be used to describe the attentional behaviours of the test-takers.

4.4.1 Demographics and Performance

Twenty-four participants were involved in Study 2. The average age of the participants involved was 15.5 years (*SD*: 0.25). The average performance of participants in their most recent school-based English, mathematics and science assessments (calculated from self-reports) was 68.04%, 60.10% and 73.10% respectively. The mean performance of these participants in Study 1B was 59.4%. In Study 2, the mean performance of these participants was 57.8% (*SD*: 5.5). Table 4.13 summarises this information.

Table 4.13 Demographic and Performance Details: Study 2

	M	SD
Age	15.5	0.3
Average Performance in English Assessments (%)	68.0	9.9
Average Performance in Mathematics Assessments (%)	60.1	14.0
Average Performance in Science Assessments (%)	73.1	11.3
Study 1B Score (%)	59.4	0.5
Study 2 Score (%)	57.8	5.5

n=24

Five tasks were included in Study 2 ('Running in Hot Weather'). Each task had an item that the test-takers had to complete. In most cases, it was a multi-part item. All tasks required test-takers to complete an SR-type item part (Part 'A'). For Tasks 2-5, test-takers also had to select simulation data to support their answer for Part A (Part 'B'). Tasks 3-5 had a CR part to the item, where participants had to explain or justify their selections for Part AB (Part 'C'). Item statistics from PISA 2015 were not publicly available for all items. However, *a priori* difficulty ratings were accessible⁵³. Item statistics were therefore calculated using this sample's performance and compared to these difficulty ratings (Table 4.14). The difficulty indices aligned well with the PISA 2015 difficulty ratings.

Table 4.14 Difficulty and Discrimination Indices for Study 2

Task (Item Part)	Difficulty	Discrimination	<i>A Priori</i> Rating
1 (A)	0.63	0.25	3
2 (AB)	0.58	0.38	4
3 (AB)	0.71	0.50	3
3 (C)	0.17	0.00	5
4 (AB)	0.69	0.63	4
4 (C)	0.58	0.75	4
5 (AB)	0.73	0.50	4
5 (C)	0.50	0.88	4

⁵³ As rated by subject matter experts using a scale of 1 to 5 with 5 being most difficult (OECD, 2017).

Test-taker performance on Part AB was calculated using a partial credit scoring system, where the maximum score was 2 (Part AB). Part C of an item, where present, was scored separately to its respective Part AB. This was marked as 'correct' or 'incorrect' (Section 3.8.1.2). A single score representing test-taker performance on all parts of an item was computed. Table 4.15 summarises the means and standard deviations of this score and also separates and summarises the performance of participants each part of a task's item. It should be noted that very few participants received 'No Credit' for Part AB of any item ($n < 4$). Given the small sample sizes involved, to facilitate comparative analyses for Items A or Items AB, Partial Credit and No Credit performances were collapsed into a single category.

Table 4.15 Item Performance by Task

Task	Total Item Score (%)		Part AB (n)		Part C (n)	
	M	SD	Partial/ No Credit	Full Credit	Correct	Incorrect
1	66.7	48.2	9	15		
2	58.3	40.8	14	10		
3	43.8	25.8	12	12	4	20
4	63.5	43.0	11	13	14	10
5	61.5	41.0	10	14	12	12

n=24

4.4.2 Research Question 3 (RQ3): What behaviours are demonstrated by test-takers when responding to items and tasks involving simulations?

4.4.2.1 RQ3a: Is time-on-task and time-per-phase related to test-taker performance?

As seen in Table 4.16, participants spent the most amount of time completing Task 4 (M : 116.0, SD : 37.0) and the least amount of time on Task 1 (M : 56.4, SD : 15.9). Given the OECD difficulty ratings for each of these tasks and their items, this is hardly surprising. The relationship between participant performance on the item (Part AB, Part C) for each task and the total time-on-task was also considered. With one exception, the amount of time participants spent completing the task was not related to participant

performance. A strong, negative and statistically significant relationship existed between participant performance on Task 1 and their total time spent on that task, $r_s = -.68$, $p < .001$. Less time spent on Task 1 was associated with better performance on the task's item. The average time spent completing items for all five tasks was 447.4s (SD : 93.9). There was no significant relationship between participant performance in Study 2 and their overall time spent completing all five tasks in the unit, $r = -.06$, $p = .80$ ($n = 24$).

Table 4.16 Mean Time-on-Task(s) by Total Item Score

Task	M	SD	r_s	p
1	56.4	15.9	-.68	<.001*
2	65.7	24.3	.15	.49
3	109.6	38.6	.22	.30
4	116.0	37.0	-.03	.89
5	99.7	26.4	.14	.51
	447.4	93.9	-.06 ^a	.80

n=24

Each task had two Times of Interest (TOIs): the 'Orientation' phase and the 'Output' phase⁵⁴. Figure 4.4 shows the mean distribution of time spent by participants in both phases for each task. For Tasks 2, 3 and 5, participants spent approximately one-third of their time-on-task in the Orientation phase. However, for Task 1, participants instead spent half of their total time in the Orientation phase. In Task 5, participants spent one-fifth of their time in the same phase, spending much more time in the Output phase as a result.

⁵⁴ The 'Orientation' phase for each task began when the task was first presented and ended when the first simulation was executed. The second TOI ('Output' phase) began when the results of the first simulation were presented and continued until the participant had navigated away from the task.

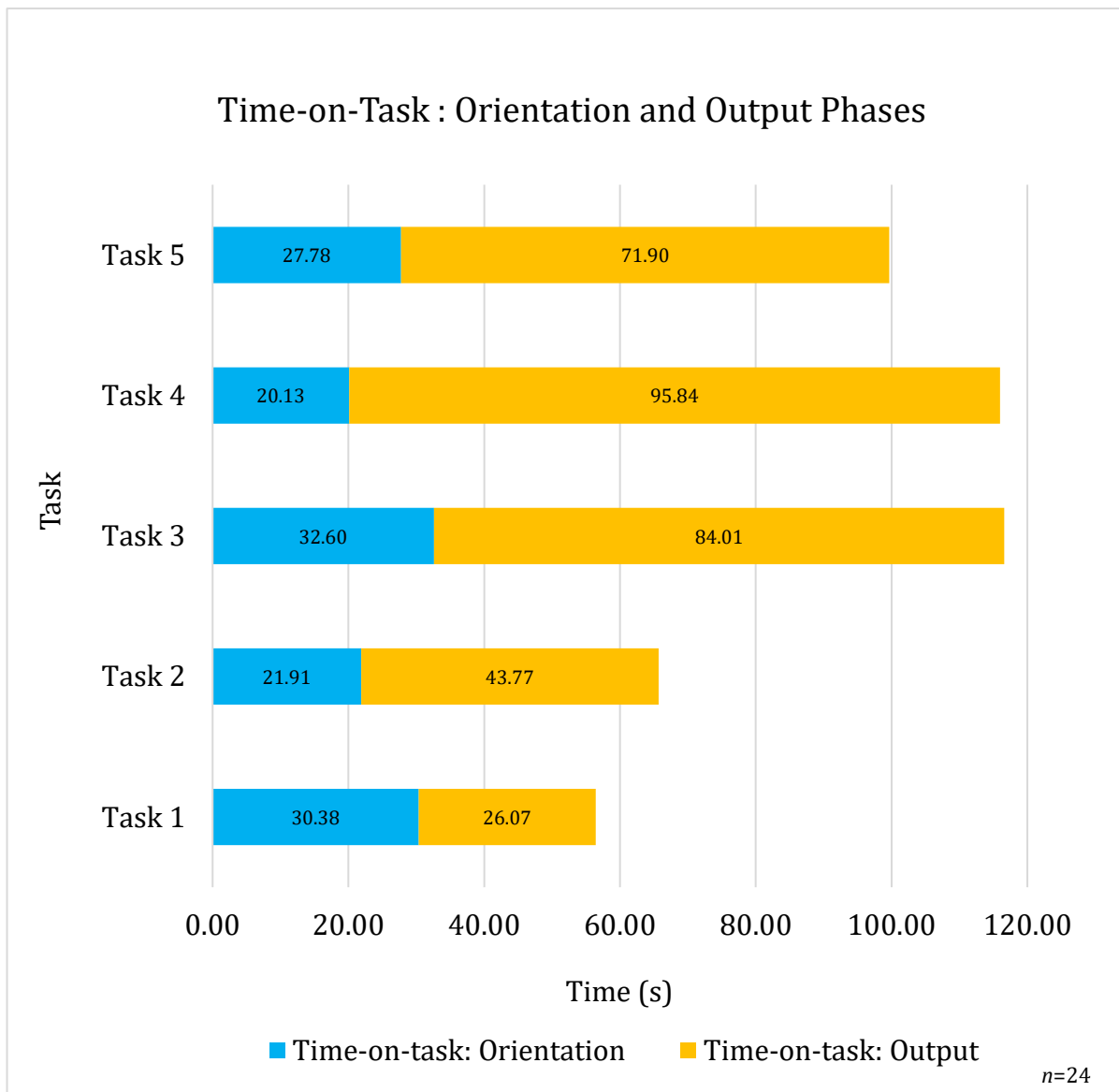


Figure 4.4 Time-on-Task for Orientation and Output phases

As illustrated by Table 4.17, total item performance (Part AB, Part C) was *not* strongly associated with time spent on either phase, with the exception of Task 1, where there was a strong, negative association between time spent in the Orientation phase of this task and scores on Task 1, $r_s = -.52$, $p = .01$. Less time in this phase of Task 1 was associated with a better total item score.

Table 4.17 Mean Time-on-Phase (per task) correlated with Item Performance (all parts)

Task	Orientation Phase		Output Phase	
	r_s	p	r_s	p
1	-.52	.01*	-.38	.65
2	-.08	.70	.20	.35
3	-.04	.87	.18 ^a	.42
4	-.26	.22	.07	.73
5	-.11	.61	.22 ^b	.31

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

a. $n=22$

b. $n=23$

4.4.2.2 RQ3b: What relationship, if any, does the number of simulations run per task have on test-taker performance?

For each task in Study 2, the participants were required to run a number of simulations to gain the information they needed to complete the task's associated item. Each task in this unit had a minimum number of simulations assigned to it by the PISA 2015 guidelines (OECD, 2015). According to Table 4.18, the median number of simulations conducted by the participants for each task generally aligned with PISA 2015's minimum requirements. Task 4 was one exception; the median number of simulations per participant for this task was 3.5 in contrast to PISA 2015's recommendation of 2 simulations. As this was a relatively small sample with a number of tied ranks, a Kendall's tau correlation coefficient was calculated to examine the relationship between the number of simulations run by a participant and their performance on Part AB of the task's item. No correlation coefficient was computed for Item 1 as all participants conducted the same number of simulations for this item. As seen in Table 4.18, there was a strong, positive relationship between the number of simulations run and participant performance in Task 2 ($r_\tau=.58$). There was also a moderate, positive relationship between these two variables for Items 3 ($r_\tau=.38$) and 4 ($r_\tau=.38$). A weak positive relationship was evident for Task 5 but this was not found to be a statistically significant difference.

Table 4.18 Relationship between Number of Simulations run and Item Performance (all parts)

Task	Minimum no. of simulations ^a	Md. Simulations	Md. Score	τ	p
2	2	2.0	50.0%	.58	.00*
3	2	2.0	50.0%	.38	.03*
4	2	3.5	100.0%	.38	.03*
5	2	2.0	75.0%	.22	.24

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

4.4.2.3 RQ3c: What attentional behaviours (*number of visits*) do test-takers exhibit when completing simulation-type items in the Orientation phase?

In the Orientation phase of each task, participants familiarised themselves with the requirements of the task and prepare to a simulation to test their hypothesis. To achieve this, participants were required to identify relevant information from the task's item on the left hand side of the screen (*Item AOI*) and then use it to design a simulation on the right-hand side (*Simulation AOI*). Areas that would house the different parts of the simulation's output were also available to view in this phase (*Relevant AOI*). The heat map⁵⁵ shown in Figure 4.5 is for Task 3 and it illustrates the focus of visual attention for *all* of the participants involved in Study 2 according to fixation counts. Unsurprisingly, this map shows that participants had a large number of fixations around the item's question stem (indicated by the red area), followed by the simulation controls (in yellow). However, the heat map suggests that participants did not allocate much of their attention to the areas that would contain the relevant information in the Output phase. For Task 3, this would be areas involving 'Sweat Volume'.

⁵⁵ It is challenging to visually compare data from participants when the dynamic content of the task makes each participant's interactions with the simulation different (Carter & Luke, 2020). To account for this, all heatmaps in this thesis are based on 'relative counts'. This is calculated by the number of fixations relative to the *total* number of fixations made by the participants during the TOI. This is most appropriate for use when the duration of time spent in a TOI can vary among participants.

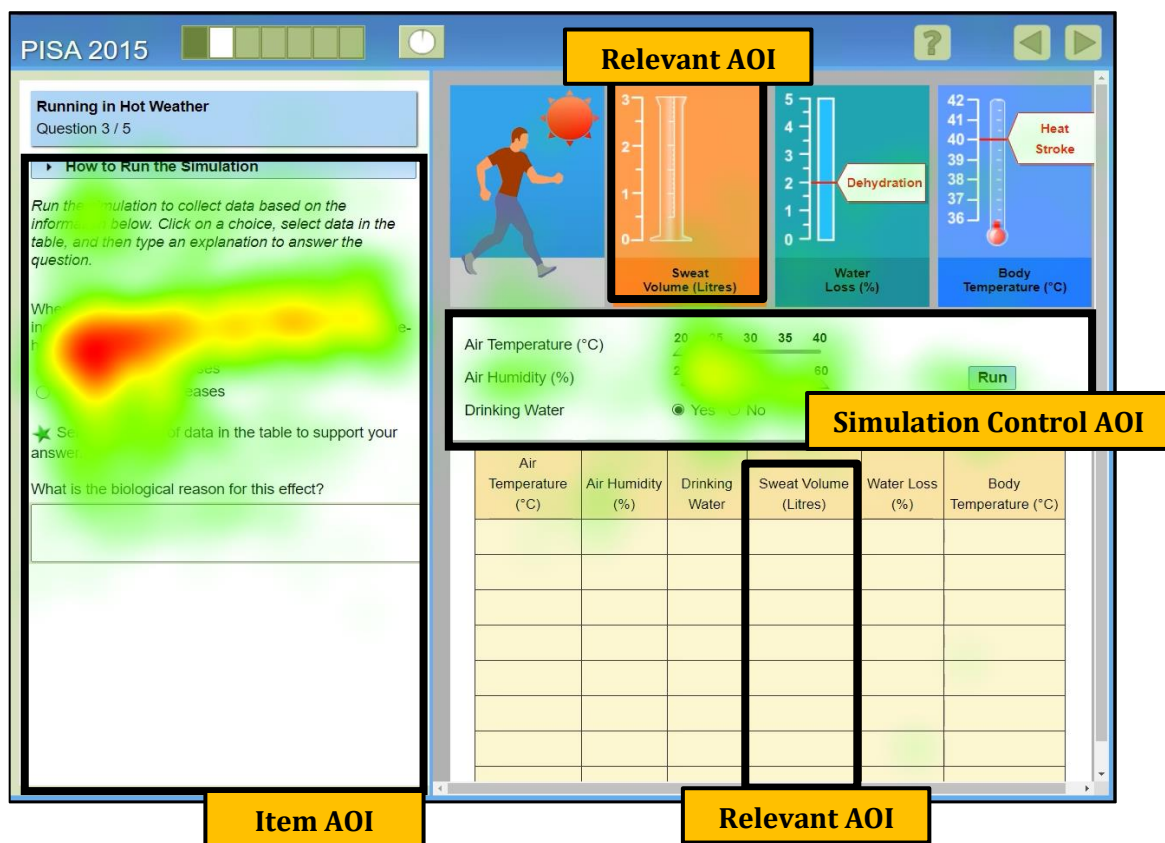


Figure 4.5 Heat map for the 'Orientation' phase of Task 3 ($n=24$)

Table 4.19 illustrates the mean number of visits by participants (how often a participant looked at, left and then returned to an AOI) on the *Item AOI*, *Simulation AOI* and *Relevant AOIs* for each task⁵⁶. Participants visited the Instructions AOI most often, indicating that they regularly looked to the task's item for guidance on what simulation to run first. Tasks 1 and 2 had a number of visits to the Simulation Control AOI. They left and then returned to this area most frequently during the Orientation phase of these tasks. For Tasks 3, 4, 5 visits to the area were lower, suggesting that when participants were fixated on this AOI they did not leave it. Participants did not appear to pay much attention

⁵⁶ While the fixation counts used to generate these heatmaps are a useful measure of attentional focus, they cannot provide much insight into participant behaviours in the Orientation phase of an item. There are no data contained in the simulation areas on the right-hand side of the screen during this phase. As a result, fixation-based metrics are not appropriate to use as participants have nothing to fixate on. Instead, visit counts can provide a clearer insight into test-taker behaviours during the Orientation phase of a task as it provides a count of how often an AOI was *returned to* (Tobii Pro, 2020). In this study therefore, it can be used as a measure of how relevant participants considered an area to be to the task's requirements.

to those areas that would eventually contain the information they needed to answer test items (Relevant AOI).

Table 4.19 Mean Number of Visits to AOIs in Orientation Phase

Task	Instructions AOI		Simulation Controls AOI		Relevant AOI	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
1	7.04	2.31	7.54	3.01	1.42	2.02
2	5.67	1.90	5.79	2.25	0.54	0.78
3	4.75	2.33	5.12	2.49	0.71	1.20
4	4.58	2.23	4.83	2.32	0.63	0.92
5	4.83	2.90	4.79	2.96	0.46	0.78

4.4.2.4 RQ3d: What attentional behaviours (*time-to-first-fixation, number of whole fixations, proportion of fixations*) do test-takers exhibit when completing simulation-type items in the Output phase?

To understand how participants interacted with the data they generated in the Output phase of each task, three eye movement measures were examined: Time-To-First-Fixation, number of fixations and proportion of fixations. As explained in Section 3.8.1.2, test-taker performance for Part AB and Part C of a particular task's item were considered and scored separately. Test-taker performance on Part AB for each task's item was used in the subsequent analyses to represent test-taker performance for a task⁵⁷.

Time-To-First-Fixation

Time-To-First-Fixation is an early measure of eye movement. It represents how long before a region of interest is fixated upon by an individual and is thus considered a measure of visual search efficiency (Carter & Luke, 2020). As in the Orientation phase, area(s) of the simulation that contained the information necessary to the successful completion of a task's item were tagged as 'relevant' in the Output phase. Moderate or weak negative correlations between performance on a task's item and time-to-first-fixation by participants on relevant information were found using (Task 2: $r_s = -0.40$, Task

⁵⁷ Part C was *not* used as a measure of test-taker performance as test-taker success with this part was not always reliant on the on the actions or behaviours of test-takers e.g. some answers could be based on prior knowledge (Item 3C).

3: $r_s = -0.38$, Task 4: $r_s = -0.15$, Task 5: $r_s = -0.03$). The less time it took for participants to fixate on relevant information, the better their performance. Only one task, Task 1, had a statistically significant correlation between item performance and time-to-first-fixation; $r = .49$. This was a *positive* relationship however, whereby the *more* time it took for participants to find relevant information, the better their performance.

Table 4.20 documents how quickly the entire sample fixated upon a relevant area of interest after their *first* simulation was executed. Given its relative simplicity, it is unsurprising that participants were most efficient at finding relevant information for completing Item 1 ($M: 442.2\text{ms}$, $SD: 416.8$). In contrast, the mean time for participants to first fixate upon a relevant area for Task 5, which required participants to integrate information from multiple variables in the simulation, was 3337.6ms ($SD: 6498.8$).

Table 4.20 Mean Times-To-First-Fixation (ms) on Relevant AOIs (Output Phase)

Task	All	
	<i>M (ms)</i>	<i>SD</i>
1	442.2	416.8
2	2994.7	8551.9
3	2194.2 ^a	5140.8
4	3337.6	6498.8
5	3988.7 ^b	8861.8

a. $n=22$

b. $n=23$

Table 4.21 represents the differences in mean times-to-first-fixations between those who received full credit for performance on Parts AB of a task item and those who did not. The descriptive statistics that form the basis for both the table and the graph indicate that for some items, those who received full credit for the item were quicker to fixate on relevant information in the simulation compared to those who received partial or no credit. For example, those who received full credit for the item in Task 2 fixated on the relevant information in the simulation in a relatively short period of time, $M: 902.7\text{ms}$ ($SD: 1703.7$). Those who received no credit or partial credit for the item took much longer to find the relevant information, $M: 4489.0\text{ms}$ ($SD: 11027.8$). However, for other tasks,

the inverse was true. This was evident in Task 4, where participants who did not receive full credit for their performance on the task were faster at attending to relevant information (M : 2824.8ms, SD : 5089.1) than those who did (M : 3771.5, SD : 7675.9).

Table 4.21 Mean Times-To-First-Fixation (ms) on Relevant Areas by Item Performance

Task	Partial/ No Credit			Full Credit		
	<i>M (ms)</i>	<i>SD</i>	<i>n</i>	<i>M (ms)</i>	<i>SD</i>	<i>n</i>
1	317.1	283.1	9	517.3	472.8	15
2	4489.0	11027.8	14	902.7	1703.7	10
3 ^a	3783.5	7430.8	10	869.8	1031.7	12
4	2824.8	5089.3	11	3771.5	7675.9	13
5	5733.4	13093.5	10	2646.5	3302.6	13

a. $n=22$

As shown in Table 4.22, Mann-Whitney U tests were conducted comparing time-to-first-fixation across the two item credit categories (Full Credit, Partial/No Credit). While there were no statistically significant differences in times-to-first-fixations for these groups, small to medium effect sizes were noted for four out of the five tasks.

Table 4.22 Time-To-First-Fixation on Relevant Information (ms) in Output Phase: Comparisons of Partial/No Credit (Group 1) and Full Credit (Group 2)

Task	<i>U</i>	<i>z</i>	<i>p</i>	<i>r</i>	Group 1	Group 2
					<i>Md</i>	<i>Md</i>
1	96.00	1.70	.10	.35	146.0	404.0
2	42.00	-1.64	.11	-.26	671.0	200.0
3 ^a	45.00	-0.99	.35	-.21	775.0	239.5
4	54.00	-1.01	.33	-.21	992.0	229.0
5 ^b	63.00	-0.12	.93	-.03	798.0	454.0

a. $n=22$

b. $n=23$

Number of Fixations

Late measures of eye movement include the number of fixations within a region of interest. It represents the total number of fixations made by a participant within a region of interest for a set period of time (Carter & Luke, 2020). For Study 2, the number of fixations in an area was used as measure of total attention. Table 4.23 outlines the mean number of fixations for the entire sample on relevant areas for each item. Task 4 had the highest mean number of fixations on relevant areas (M : 43.1, SD : 28.6). This is hardly surprising given that Task 4 required the highest number of simulations, and thus required the most attention. However, when the mean number of fixations on relevant areas was broken down by performance on Part AB, some variations in the number of fixations conducted by participants were noted. For example, the mean number of fixations on relevant areas for Task 1 was nearly identical across all performance categories (Partial/ No Credit; M : 16.3, Full Credit; M : 14.1). For the remaining items, high performers generally had *more* fixations on relevant areas than low performers and moderate performers. With one exception (Task 4), they appeared to pay *more* attention to relevant areas.

Table 4.23 Mean Number of Fixations on Relevant Information (Output Phase)

Task	All		Partial/ No Credit			Full Credit		
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>
1	14.9	10.3	16.3	11.2	9	14.1	10.0	15
2	30.7	34.2	24.8	39.9	14	38.9	23.6	10
3 ^a	31.0	22.0	24.5	16.5	10	36.4	25.1	12
4	43.1	28.6	47.6	35.3	11	39.2	22.2	13
5 ^b	22.9	19.3	21.9	20.4	10	23.6	19.2	13

a. $n=22$

b. $n=23$

Mann-Whitney U tests were conducted comparing the mean number of fixations within each item credit category (Full Credit, Partial/No Credit). Table 4.24 summarises the outcomes of these tests. For four of the five tasks, there were no statistically significant

differences in the number of fixations on relevant areas according to performance category for these groups. However, small to medium effect sizes were noted for Tasks 1, 3 and 5. For Task 2, those who received full credit for their performance had significantly more fixations within the relevant areas of interest than those who did not.

Table 4.24 Number of Fixations on Relevant Information in Output Phase: Comparisons of Partial/No Credit (Group 1) and Full Credit (Group 2)

Task	<i>U</i>	<i>z</i>	<i>p</i>	<i>r</i>	Group 1	Group 2
					<i>Md</i>	<i>Md</i>
1	57.00	-0.63	.56	-.13	15.0	13.0
2	111.50	2.43	.01*	.50	11.5	35.5
3 ^a	83.00	1.52	.14	.32	23.5	32.5
4	68.50	-0.17	.86	-.03	32.0	32.0
5 ^b	78.50	0.50	.62	.10	12.5	23.0

* $p=0.05$

a. $n=22$

b. $n=23$

Proportion of Fixations

The proportion of fixations allocated by participants to different parts of the simulation area were also considered for Study 2. The heatmap contained in Figure 4.6 illustrates how the fixation counts were distributed during the Output phase of Task 3. Participants had the highest number of fixation counts within the *Item AOI* which is unsurprising. High fixation counts were also evident for the simulation controls. To determine if participant behaviour aligned with the features of the expert-novice paradigm that underlies the CTMA (Kirschner et al., 2016), comparisons on the proportion of fixations on *Relevant* and *Irrelevant AOIs* were conducted for the whole sample and for those who had different categories of performances.

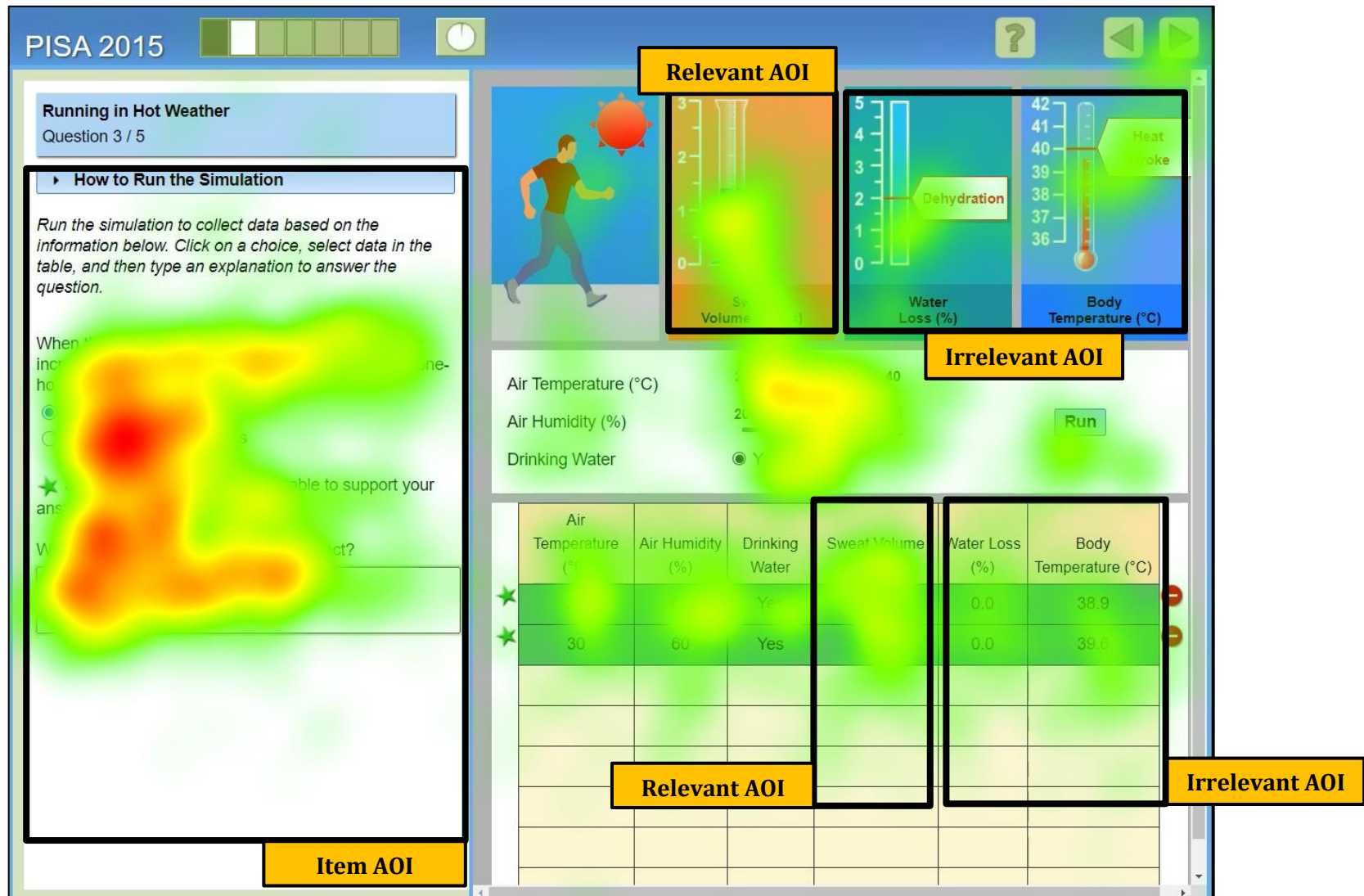


Figure 4.6 Relative count of fixations for Task 3

The mean proportion of fixations on the relevant and irrelevant areas of the simulation output are available in Table 4.25. For four of the five items, participants spent more time fixating on relevant areas of the simulation rather than irrelevant areas, as seen in Item 4 (Relevant: M : 0.25, SD : 0.1; Irrelevant: M : 0.07, SD : 0.1). However, this did not appear to be the case for Task 1, where participants' distribution of fixations between relevant and irrelevant simulation areas was relatively similar. Comparisons on the proportion of fixations on relevant and irrelevant information using Wilcoxon rank-sum tests indicated that for all tasks, with the exception of Task 1, participants allocated significantly more of their fixations to relevant areas of the simulation output. Moderate to large effect sizes were noted for each comparison.

Table 4.25 Mean Proportion of Fixations on Relevant and Irrelevant Areas (Output Phase)

Task	Relevant		Irrelevant		Wilcoxon Rank-Sum			
	M (SD)	Md	M (SD)	Md	W	z	p	r
1	0.18 (0.1)	0.2	0.22 (0.1)	0.2	185.00	1.90	.06	.39
2	0.20 (0.1)	0.2	0.03 (0.0)	0.0	0.00	-4.29	<.001*	.88
3 ^a	0.18 (0.1)	0.2	0.11 (0.1)	0.1	34.00	-3.00	.03*	.64
4	0.25 (0.1)	0.3	0.07 (0.1)	0.1	0.00	-4.29	<.001*	.88
5 ^b	0.21 (0.1)	0.2	0.11 (0.1)	0.1	28.00	-3.35	.001*	.70

* $p=0.05$

a. $n=22$

b. $n=23$

To determine if the proportion of fixations on relevant information differed according to task performance, a Mann-Whitney U test was conducted for each item credit category (Full Credit, Partial/No Credit). Table 4.26 summarises the outcomes of these tests. As a second comparison was conducted for each task, a Bonferroni correction was applied to reduce the risk of Type I errors (Von der Malsburg & Angele, 2017), reducing the p -value to 0.025. For Tasks 1-3, the moderate to large effect sizes calculated indicate that test-takers who received full credit had a higher proportion of fixations on

relevant areas than those who did not receive full credit. This result was statistically significant for Tasks 2 and 3. For Tasks 4 and 5, there was no significant or practical difference between the proportion of fixations on relevant areas by either credit category.

Table 4.26 Proportion of Fixations on Relevant Information in Output Phase: Comparisons of Partial/No Credit (Group 1) and Full Credit (Group 2)

Task	<i>W</i>	<i>z</i>	<i>p</i>	<i>r</i>	Group 1	Group 2
					<i>Md</i>	<i>Md</i>
1	44.00	-1.53	.17	-.31	0.14	0.21
2	106.00	2.11	.04*	.43	0.14	0.27
3 ^a	100.00	2.64	.01*	.56	0.12	0.21
4	71.00	-0.03	.98	-.01	0.26	0.24
5 ^b	67.00	-0.18	.87	-.04	0.20	0.20

* $p=0.025$

a. $n=22$

b. $n=23$

4.4.3 Summary: Study 2

Study 2 aimed to describe how test-takers engage with simulation-type items. There was no statistically significant relationship between participant performance in Study 2 and the overall time spent completing all five tasks in the unit. Furthermore, task performance was not strongly associated with the amount of time spent on either the Orientation or Output phase of a task (with the exception of Task 1) (RQ3a). In contrast, the number of simulations run by a test-taker did appear, for the most part, to be positively associated with test-taker performance (RQ3b). In terms of attention, minor changes to participants' eye movements in the Orientation phase of a task were noted. Descriptive statistics indicated that test-takers spent less time attending to simulation controls as they became more familiar with the tasks (RQ3c).

Eye movement data collected during the Output phase of a task indicated a weak to moderate relationship between task performance and time-to-first-fixation by participants on relevant information (with the exception of Task 5). Furthermore, it

emerged that participants allocated significantly more of their fixations to relevant areas of the simulation output than the irrelevant areas. In examining the differences between those who received full and partial/no credit for an item, performance was associated with shorter times-to-first-fixations. While the differences between these groups were not statistically significant, small to medium effect sizes were noted. No statistically significant differences in the number of fixations on relevant areas according to performance category were found either. However, once again, small to medium effect sizes were noted. Those who received full credit for a task had a higher proportion of fixations on relevant areas than those who did not receive full credit but this was not a consistent finding across all tasks (RQ3d).

4.5 Study 3

For Study 3, qualitative data from a cued-Retrospective Think-Aloud (c-RTA) protocol were collected from a small sub-group of participants who were involved in Study 1B and Study 2 ($n=12$). Participants self-selected to be involved in Study 3. For this study, participants watched a replay of their eye movements from Study 1B and Study 2 and were asked to state out loud to the researcher what they were thinking at different points of the replay video. While Study 1B and Study 2 gathered quantitative data to determine how test-takers engage with items in TBAs (RQ4), Study 3 addressed the same aim using a qualitative approach.

4.5.1 Demographics

Five participants completed the static version of the TBA used in Study 1B and seven participants completed the dynamic version⁵⁸. Two-thirds of the participants had high levels of prior scientific knowledge ($n=8$) and enjoyed Science as a school subject. In relation to Study 1B, four participants achieved a 'low' score ($<55\%$), seven achieved a 'moderate' score ($>55\%$) and one participant was classified as a 'high' scorer ($>75\%$). For Study 2, there was one participant with a 'low' score, five with 'moderate' scores and six participants attained a 'high' score. An overview of the relevant background variables for each interviewee can be seen in Table 4.27.

⁵⁸ Consent for involvement in Study 3 was obtained at the start of the research study. As a result, it was not possible to plan in advance what condition the participants involved were assigned to or what their background variables were.

Table 4.27 Profile of Interviewees for Study 3

	Condition	Average Science Score (Level)*	Enjoyment of Science**	Study 1B: Score	Study 2: Score
Interviewee 1	Static	72% (M)	2	73% (M)	81% (H)
Interviewee 2	Dynamic	80% (H)	3	40% (L)	6% (L)
Interviewee 3	Static	64% (M)	3	100% (H)	81% (H)
Interviewee 4	Dynamic	68% (M)	3	67% (M)	63% (M)
Interviewee 5	Dynamic	89% (H)	3	67% (M)	63% (M)
Interviewee 6	Dynamic	87% (H)	3	67% (M)	81% (H)
Interviewee 7	Dynamic	72% (M)	3	53% (L)	81% (H)
Interviewee 8	Dynamic	87% (H)	3	67% (M)	100% (H)
Interviewee 9	Static	67% (M)	3	60% (M)	69% (M)
Interviewee 10	Dynamic	75% (H)	3	53% (L)	63% (M)
Interviewee 11	Static	65% (M)	3	47% (L)	56% (M)
Interviewee 12	Static	83% (H)	3	73% (M)	88% (H)

* Average Science Score (Level) refers to the average performance of participants in their school-based tests/assessments in the subject (self-report). This was then used as a proxy measure for their level of prior science knowledge which was classified as Low (L; <54%), Moderate (M; >55%) or High (H; >75%).

** This is the median score the participant received on the 'Interest in Science' scale. Scores ranged from 1-4, with higher scores indicating greater enjoyment of science-related activities.

The data collected from these participants were analysed using Braun and Clarke's (2006) six-step framework for thematic analysis. *NVivo 12* software (QSR International, 2020) programme was utilised to facilitate the process. Figure 4.7 summarises the final thematic framework that was constructed based on the qualitative data. Two levels of themes are present in the thematic framework as per Braun and Clarke's (2006) definitions. The first of these, semantic themes, relate to the 'explicit or

surface meanings of the data where the analyst is not looking for anything beyond what a participant has said' (p. 84). In contrast, latent themes 'examine the underlying ideas, assumptions and conceptualisations... informing the semantic content of the data' (Braun & Clarke, 2006, p. 84). Both theme types were identified during the analysis process. As seen in Figure 4.8, one semantic theme was identified in the data where participants offered their opinions of and recommendations for the future of online assessments ('*Feedback*'). This theme was fundamentally different to the three other latent themes that were also constructed. These three themes ('*Familiarisation*', '*Sense-making*' and '*Making Decisions*') did more than summarise the participants' responses on a topic. Instead, these themes attempted to capture the nature of the participants' interactions with online testing environments. These themes will now be discussed in turn, beginning first with the identified latent themes.

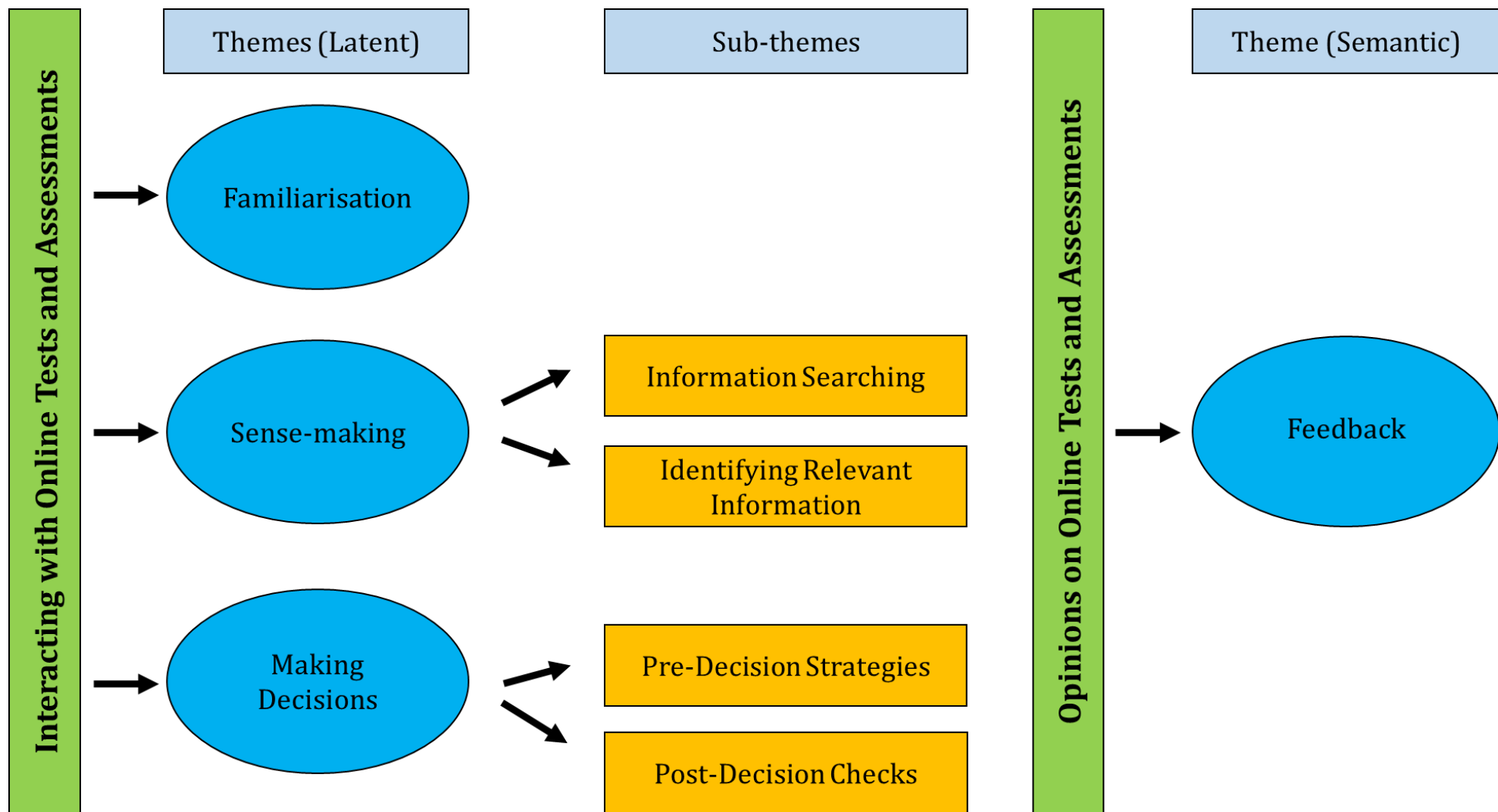


Figure 4.7 Thematic frame representing principal themes and subthemes

4.5.2 Familiarisation

The first latent theme, *Familiarisation*, reflects the means through which the participants orientated themselves to the requirements of the online testing environment and the overall value they placed on this process. All of the participants mentioned that it was important to first '*figure out what to do*' (Interviewee 2) when they logged on to the online testing platform. During this time, the participants asserted that any confusion regarding the overall layout of the system needed to be overcome as soon as possible e.g. how to progress, how to select an answer. Many of the participants noted that the volume of information on the screens during these practice items required them to actively pause and search for the spatial *position* of key elements (e.g. question, response options) to use as 'checkpoints' *before* cognitively engaging with them.

I was just looking all around the screen cos I was trying to find that actual question.

Interviewee 4

It wasn't too bad. I was just kind of overwhelmed with the video being there as well as the question. I took a minute to get used to it 'cos there's a lot on screen.

Interviewee 8

All of the participants noted the practice items for Study 1B and Study 2 were necessary and valuable. Yet, the familiarisation process appeared to be repeated each time a new item type was encountered. For example, two of the FR items in Study 1B required participants to 'drag-and-drop' stimuli into a particular order. However, when reviewing their eye movements for the first of these FR items, some participants admitted that they ignored the instructions explaining this requirement as they did not have a similar appearance to the instructions found in other items (they were not italicised). Instead, they immediately engaged with the test stimuli to understand how the question could be completed before engaging with the actual content of the test item.

Researcher: And what did you think of having the 'A', 'B' and 'C' cards there? Did you understand immediately what to do?

Interviewee 6: I wasn't entirely sure at first.

Researcher: Mm hmm... So why did you hold the 'B' card and then let it go?

Interviewee 6: Yes. Just to see what would happen.

Some participants noted that their understanding of FR type items was aided by their experiences with other online environments e.g. dragging and dropping browser tabs (Interviewee 9), playing games on the DS (gaming console; Interviewee 7). Others realised that they failed to transfer their knowledge from non-assessment online platforms (Interviewee 2). Regardless of their prior knowledge of online environments, becoming more familiar with the testing platform and the test items was appreciated by the participants. This process of familiarisation was considered necessary to the management of their overall cognitive load within the assessment process. Becoming familiar with the items allowed them to develop more efficient information search strategies so that they could focus on answering the questions. This was particularly relevant in Study 2B, where highly novel item types were presented to the participants. The practice item for this study, therefore, was highly valued as it gave them more time to 'get used to' this item type. Participants seemed to be aware of a 'practice effect' of online tests and how it can be leveraged to support their performance in later items.

Um, yeah. But after the third question, I kind of basically knew exactly what to do and where things would pop up.

Interviewee 2

Like, initially I thought, like, 'oh wow, that's a lot of things on the screen'. But by this question [STUDY 2, QUESTION 3], it was more manageable then.

Interviewee 9

4.5.3 Sense-Making

The second latent theme, *Sense-Making*, captures the thoughts and behaviours participants engaged in when attempting to sort and use the information presented to

them. Two distinct approaches to the sense-making process were identified through data analysis: *Information Gathering* and *Identifying Relevant Information*.

4.5.3.1 Information Gathering

When the participants encountered a test item, they, for a relatively brief period of time at least, engaged in a general visual search. This general search was different from what was described in the *Familiarisation* process as the participants now took into consideration the *content* of these elements. The visual stimuli (videos, text and images) acted as an important reference point to guide participants in their efforts to understand the test item's content. However, different searching techniques were employed depending on the type of visual stimuli presented. For the participants in the Dynamic condition, the videos were generally ignored after the first item of any unit. The video was always played for the first item and then occasionally in later items if participants wished to double check something. Furthermore, the visual elements of the videos were largely ignored when they were first played. Instead, the seven participants in the Dynamic condition listened to the audio narration while they read the test item and/or response options. Some participants in the Animated condition were aware that this use of the audio narration was a key aspect in their information search strategy with Interviewee 6 noting that '*... the first time I played it, I was mainly listening to the video*'. Others, even though they had similar background variables (see Table 4.27), were not at all aware of this behaviour – it was only when they reviewed their eye movements they realised this had occurred, as demonstrated in the exchange below.

Researcher: We can see it in your eye movements that for the vast majority of the time while the video is playing...

Interviewee 5: I'm actually reading instead of looking at the video. Didn't realise that.

Those in the static condition acknowledged that the way in which information was presented to them on-screen was similar to that of a traditional 'paper-and-pencil' test, despite this information being presented to them on a horizontal rather than a vertical plane (e.g. Interviewee 3, Interviewee 9). Interestingly, some participants in the static condition felt that the images accompanying the text were not always useful. Some noted

that the images in the Power Plant and Groundwater Extraction units were the *only* beneficial ones as they were diagrams rather than pictures. Diagrams were highly valued by the participants in the static condition as they gave '*the gist of what you're going to be working on*' (e.g. Interviewee 9). Indeed, as indicated by more than half of the participants in the dynamic condition, static diagrams were so desirable in the TBA that these participants often created their own in the final two units of the TBA. Interviewee 8, who had a high level of prior scientific knowledge but who achieved a 'moderate' score in Study 1B, explained that pausing the videos in the Dynamic condition (and thus, unintentionally creating a static item) was a more efficient way for them to gather information when the videos were explaining diagrams.

For some test-takers, the layout of the test item hampered the efficiency of their search strategy. In the third item of the Meteor Unit (involving a drag-and-drop response action), participants had to refer to a picture depicting three different sized craters to answer two test items. This was housed in the left hand side of the screen. The contextual information for the unit was on the right hand side of the screen. Interviewee 8 found this particularly confusing, stating that '*...because the video was bigger than actual picture needed to answer the question... it took me a while to find the question and the instructions*'. All of the participants, regardless of their condition, noted some level of difficulty in their interactions with the final two items of the Meteoroids and Craters unit. The layout of simulation-type items was also considered to be confusing by at least three-quarters of the participants.

4.5.3.2 Identifying Relevant Information

In general, once the participants had gained an understanding of the overall position and content of the test item's elements, the process of identifying *relevant* information began. This was any information that they believed would help them to complete the test item. A number of test-taking strategies supported this more focused search for information. Some appeared to have 'transferred' over from paper tests and others were specific to the testing environment or condition. Unsurprisingly, most participants admitted that they spent some time trying to recall what '*they already knew from Science class*' about a particular topic (Interviewee 7). Furthermore, those in the static condition attempted to find and match key words from the stimulus text and the question stem. Many participants in this condition noted that this was a standard test-

taking strategy that they felt comfortable using in an online environment. They did this consistently in every item, even for those items where the stimulus text was the same as the previous item.

I'm just trying to find a keyword that I can find in both the text and answer. I always do that in every question.

Interviewee 9

For those in the Dynamic condition however, there was no opportunity to do this. Instead, these participants had to 'listen out' for the keywords in the audio narration. To speed up this process, all of the seven participants interviewed noted that they skipped through the video listening out for a key word or visual cue. For some, the use of videos as a presentation format was frustrating as '*you don't get the information immediately*' (Interviewee 8). Others felt justified in not playing the video more than once in a unit as they '*didn't need all of it again*' (Interviewee 8) or '*remembered the content pretty well from the last two questions*' (Interviewee 10). Although the videos often slowed down their search for relevant information, they did appear to provide other contextual information that students considered relevant. In attempting to determine which energy conversions occurred in the Power Plant unit, Interviewee 5 said that they '*remembered the lights lighting up*' in the video, thus making them more confident that at least one of the energy forms involved was electrical. Similarly, Interviewee 8 highlighted that skipping to the end of the video allowed them to watch how '*the electricity came on after the water made the turbine move*'. This gave them the information they needed to complete one of the test items in the Blue Power Plant unit. This animated representation of energy conversions in the Blue Power Plant unit may have allowed participants in the dynamic condition to more easily identify the relevant types of energy involved compared to those in the static condition.

For the simulation-type items (Study 2), some participants admitted to identifying in advance what areas of the simulation output they should attend to *before* they ran the simulation (e.g. Interviewee 4, Interviewee 5, Interviewee 9). As illustrated by Interviewee 4, the amount of information available in the simulation-type items required a more efficient search for relevant information based on the content of the test question. However, not all participants employed such a focussed approach to searching for the

relevant information they, with others waiting until *after* the simulation had been run to look for the relevant information needed (e.g. Interviewee 1, Interviewee 6, Interviewee 11).

Interviewee 4: I took a guess before I ran a simulation. Then I could look at all the stuff then afterwards to see if I was right or wrong.

Researcher: OK, so you actually kind of had a hypothesis going in?

Interviewee 4: Yeah. I needed it to help me know what to look at cos a lot of numbers popped up after the simulation.

Participants had to sort through a large amount of information to identify the relevant information they needed to complete the simulation-type test items in Study 2B. In attempting to identify this information, personal preferences seemed to play some role. After every simulation, a 'banner' would appear at the top of the screen highlighting if the runner was at risk of heat stroke or dehydration. This banner would disappear when another simulation was run. Every simulation was recorded in a table (although the presence of heat stroke or dehydration was not identified in the table). For some participants, the table was more useful in identifying the outcomes of their simulations. For others, the images on top with the banners were more helpful.

I just looked at the images on the top, the images on the top, and then I tried to remember which one was right... I barely ever looked at the table.

Interviewee 5

I thought the bits up top were a bit useless. The table was more useful in deciding the answer cos you had a record.

Interviewee 7

4.5.4 Making Decisions

The third latent theme, *Making Decisions*, represents the decision making process undertaken by the participants as they completed each test item. Two key stages to this process were recognised: *Pre-Decision Strategies* and *Post-Decision Checks*. The first of

these embodies *how* the participants came to their final decision based on the information they had previously deemed relevant. The second represents the final interactions the participants had with an item before moving onto the next one.

4.5.4.1 Pre-Decision Strategies

When reviewing their eye movements, the participants recalled their thoughts in selecting or constructing their final response to an item with relative ease. In making a final decision on an answer for a test item, the participants did admit to some guessing behaviour if they were unsure (e.g. Interviewee 2). However, for multi-part questions, such as those that needed participants to select two words to complete a sentence (e.g. Power Plant Unit, Item 3), this uncertainty was much easier to manage. If participants knew the answer to one part of the item, they answered that part first and then considered the other part of the item; *'I knew that it was definitely electrical for part 2 so I said I'd start with that'* (Interviewee 4). The participants acknowledged that this approach of *'start(ing) with the answer you are more confident of'* (Interviewee 7) was one they would employ in a standard pen-and-paper exam. However, it was much easier to use this strategy in an online exam.

It's just two clicks. It doesn't... It's kind of quicker than just rubbing something out and stuff. It's no big deal if you change your answer or just put down a placeholder in an online exam.

Interviewee 9

Other strategies to support their final decision were also described by the participants. For SR items like MCQs, the participants often 'eliminated' the possible response options one-by-one, even when they were confident of their answer. Interviewee 8 admitted that they knew immediately that three of the options could be eliminated but they *'needed to read it twice to make sure'*. This preference for 'double checking' information before making a final decision was evident regardless of item type. In Study 2A, participants consistently re-read test items and options *'just in case they [sic] missed something'* (Interviewee 5). In Study 2, the students had to generate their own information to answer the test items. The decision to run a particular simulation was, for most participants at least, carefully considered. Prior to running the simulations needed

for such information, the participants spent a large amount of time ‘double checking’ their work, with Interviewee 6 explaining that they ‘*had to see if they were correct... if it was set properly*’. Other participants did not pay much heed to preparing these simulations, with the decision to run a particular simulation being relatively unplanned e.g. Interviewee 1. Participants’ approaches to this did not appear to have any impact on their performance in Study 2.

4.5.4.2 Post-Decision Checks

Deciding upon a particular answer or response option did not signify the completion of a test item. Analysis of the qualitative data indicated that the time *after* making a decision on their final response to an item but *before* moving onto a new item was distinguished by a number of key behaviours among the participants. The participants reported that they spent some time checking the item one last time before moving onto the next test item. The behaviours associated with these ‘post-decision checks’ were very similar to those that constituted the pre-decision strategies. For example, after completing an item many participants spent some time ‘*double checking*’ their answers one last time (e.g. Interviewee 1, Interviewee 11). This occurred even when the participant had been confident in their final decision. When queried further on this, some participants noted that they would ‘*always do this in a test*’ (Interviewee 12) and were just transferring previously taught test-taking strategies to the online environment. However, at least half of the interviewed participants indicated that this interaction with a test item was a new experience for them that was prompted by the online environment.

Interviewee 3: Like, if I was in an exam, I usually just go over something [sic] once at the end so that I have enough time during the exam. I wouldn't like double check it straight away.

Researcher: So were you more likely to double check it on the computer?

Interviewee 3: Yeah, that's the reason I did good here I think.

The testing system did not allow the participants to review their answers before submission. While the participants were aware of this from the outset, this did not seem to be a factor that contributed to the occurrence of these post-decision

checks as no participant mentioned it. Furthermore, it was observed that the participants rarely changed their answers during the post-decision time period, thus suggesting that it was not likely that uncertainty over their answer prompted them to double check their work again. Instead, the online environment itself seems to have naturally encouraged the participants to do some post-decision checks. According to the participants, online testing environments were considered more legible than traditional paper-and-pencil approaches. Interviewee 11 noted that *'it's easier to see and spot stuff online than written down in your own handwriting I think'*. This opinion was supported by other participants too.

It's just... easier [to double check something on a screen]. Less confusing. Everything pops out at you.

Interviewee 10

4.5.5 Feedback

The final theme attempts to summarise the participants' feedback on the online test they had just completed and their view on online tests and assessments in general. In revealing their opinions, the participants offered a number of recommendations for the design and use of online tests and assessments. For example, participants had a clear preference for online exams compared to traditional pencil-and-paper tests. However, there appeared to be some conditions attached to this preference. Participants were predominately in favour of online tests for subjects that required them to generate a large amount of text in their responses to test items. Online tests would allow them to type instead of handwriting the answers. This was preferable as typed text was considered to be *'neater, quicker and easier'* (Interviewee 10). However, at least five of the interviewed participants recognised that their own typing skills would need to be addressed before they would be comfortable with the introduction of online tests for post-primary schools.

Researcher: So you're faster at writing than typing, are you?

Interviewee 12: Yeah, I can get my ideas down quicker. I'd need to learn to type properly to be happy with an online exam for the Leaving [Certificate].

Furthermore, many participants recommended that high-stakes exams for some subjects be excluded from online platforms. When discussing their thoughts for future online exams, the participants identified some subjects that, in their opinion, would be ill-suited to an online platform. These included geography, engineering and mathematics. Most of the participants indicated that the activities required of them in an exam for these subjects are difficult to do on a screen e.g. drawing diagrams, writing formulae etc. As a result, they recommended that online exams for these subjects not be considered.

Um, maths... it's just really practical and you have to write formulas down. . .And in geography, you have to draw loads of diagrams.

Interviewee 7

I don't know. It's just there's a lot of, you know... I feel like... Let's say now English would be better to, you know, do online because... Just because with maths all the numbers and equations and stuff. English or history would be OK to do because they have a lot of typing and stuff.

Interviewee 10

Uhm... maybe not woodwork? Because you have to do some sketching.

Interviewee 11

In relation to the actual design of online tests, the participants did provide some interesting insights that could inform the design of future online assessments. For example, it appeared that there was no real preference for one item type (e.g. SR, FR, CR or simulation-type items) over another. In fact, one participant noted that they *'liked the variety'* (Interviewee 10). Interestingly, two participants from the dynamic condition noted that they would have preferred to have seen text-based stimuli rather than the audio-visual stimuli they had experienced. Interviewee 4 felt that for the majority of the videos *'the picture was enough at the end'*. Interviewee 2 argued that the absence of text to refer to made some of the items *'really hard'*. Other participants did not note anything of significance in relation to the use of video-based or text-

image stimuli in the test. In contrast, participants did make an effort to note that, regardless of an item's type or its content, careful consideration should be afforded to how an item looks on a screen. Half of the participants recommended that certain aesthetics should be adhered to when designing an online screen to make it easier to interact with the test platform. These design recommendations usually related to the use of *'specific font types to indicate different things'* (Interviewee 5). Interviewee 3 recommended that *'questions should be in a different font and bold so that you can tell what's a question and what's just random'*. Another participant suggested that having blank sections and spacing between elements is important to prevent students from feeling *'overwhelmed'* (Interviewee 8).

Other general recommendations for the overall design of an online test were also highlighted by the participants. Two of the participants said that they felt reassured by the system's warnings if they had not answered a question properly or forgotten something. Yet, despite this, the participants did suggest that more navigational freedom e.g. being able to skip questions and then return to them, was needed in online exams, particularly in comparison to the test they had just completed.

You don't really have an option of skipping anything online but you might want to do some parts first. You need to be able to skip to them.

Interviewee 12

I knew where everything was and normal tests... it's happened before where I missed an entire page!

Interviewee 5

4.5.6 Summary: Study 3

The data collected from the cognitive interviews conducted in Study 3 provides important information around the nature of test-takers' interactions when attempting to complete a TBA. Test-takers indicated that they spend some of the testing time familiarising themselves with the TBA and the different item types contained within. They then 'make sense' of the item using information searching strategies to identify relevant information. Based on the data gathered, there is evidence in this study to

support the argument that that the multimedia stimulus used in an item had an impact on the information searching strategies employed. For SR and CR items, test-taker decisions on their responses comprised of two stages. The first of these involved the deployment of pre-decision strategies e.g. completing item parts in a non-sequential manner. After a decision had been made, test-takers appeared to engage in a number of post-decision checks to confirm their final selection. Participants also gave generalised feedback and recommendations on the future use of TBAs for second-level students.

4.6 Summary

This chapter presented the results regarding the impact of multimedia stimuli on test-taker performance and behaviour (Study 1A, 1B). The results of the current study showed that multimedia stimulus had no systematic impact on overall test-taker performance in a TBA. However, key item statistics (difficulty, discrimination) suggested that the items performed differently across conditions. Based on the effect sizes calculated, eye-movement data further indicated that the use of static or dynamic stimuli affected how test-takers interacted with the interaction space of an item. Analysis from data collected in Study 2 revealed that the attentional allocation behaviours of test-takers became more directed and efficient as they became more familiar with simulation-type items. Differences in attentional behaviours between those who received full credit or partial/no credit for an item were noted. While these did not always reach statistical significance, small to medium effect sizes were regularly noted. The cued-Retrospective Think Aloud (c-RTA) conducted for Study 3 highlighted some important aspects of test-takers' behaviours in online testing environments. The next chapter will provide a discussion of how these findings might be used to guide decision making around the use of different items types in TBA more generally and how future studies can be designed to address what still remains to be understood about their impact on test performance.

Chapter 5

Discussion and Conclusions

5.1 Introduction

This research investigated the impact of multimedia stimuli on test-taker performance and attentional behaviour in Technology Based Assessments (TBAs). It also examined how test-takers engage with simulation-type items. Test score, eye movement and cued-Retrospective Think Aloud data were collected from Irish post-primary school students to better understand these issues. This final chapter summarises and contextualises the major findings of this thesis in relation to previous research and literature. In particular, it discusses how the use of dynamic and static stimuli affected test-taker behaviour without any apparent corresponding effect on performance. The chapter also summarises some of the key attentional behaviours exhibited by test-takers while completing simulation-type items. Recommendations for policy, practice and future research in relation to TBAs are discussed

5.2 Summary of Research

TBAs use items that employ a broad array of dynamic or static stimuli (e.g. animations, text-image, simulations) and response mechanics. Although it is assumed that these features can make TBAs more effective, their impact on test-taker performance and behaviour had yet to be fully determined (e.g. Bryant, 2017). Using a mixed methods approach, this research aimed to investigate the extent to which the use of different multimedia stimuli could affect test-taker performance on items in a TBA. To address this, an experiment was first conducted with 251 Irish post-primary students using an animated and text-image version of the same TBA of scientific literacy. Eye movement ($n=32$) and interview data ($n=12$) were also collected to determine how these multimedia stimuli affected test-taker attentional behaviour in relation to the interaction space⁵⁹ of an item. A second study involving 24 test-takers completing a series of simulation-type items while monitored by an eye-tracker was also conducted. Interview data were also gathered here.

⁵⁹ That space where the test-taker's actions and responses are recorded (Russell, 2016).

5.3 Key Findings

5.3.1 *Static and Dynamic Objects as Item Stimuli*

5.3.1.1 Impact on test-taker performance and item functioning

Answering test items in a TBA is a complex retrieval task that is affected by the content, readability, and layout of the item's interaction space and stimulus (Russell, 2016). Based on the data collected in Study 1A from 251 Irish post-primary students, the type of multimedia stimulus used in an item does not appear to have a direct effect on test-takers' performance in a TBA. However, prior knowledge did appear to interact with stimulus type to affect test-taker performance. Specifically, test-takers with low levels of prior knowledge performed better when static stimuli were used, whilst those with high levels of prior knowledge performed better in the dynamic condition. This effect was small, and, although not significant in this small-scale study, reflects previous findings (e.g. Malone & Brünken, 2013; Wu et al., 2015) and could have some practical importance.

Using the theoretical frameworks proposed by Mayer (2014) and Kirschner et al. (2016) as an explanatory guide, students with high levels of prior knowledge may have performed better in the dynamic condition because they had sufficient cognitive resources to process the *extra* information contained within the narrated animations. The differing multimedia stimuli allowed for different levels of cognitive assessment load (Kirschner et al., 2016). Dynamic stimuli, such as animations, allow viewers to perceive moving phenomena as they would in the real world e.g. magma flow, energy conversions. Participants interviewed in Study 3 noted that the dynamic nature of the animations allowed them to better understand and recall the energy conversions being examined in the 'Blue Power Plant' unit. The dynamic multimedia stimuli appeared to provide test-takers with additional 'cues' not available with static stimuli. Students with low levels of prior knowledge appeared unable to manage or interpret this information (also known as germane assessment load; see Figure 2.10), hence their relatively better performance in the static condition. It should be acknowledged that Wu et al. (2010) found that test-takers of a similar age to those in this study with high levels of prior knowledge performed better with static stimuli. Those with lower levels performed better with dynamic stimuli. Subtle differences in instrumentation may have contributed to the divergence in findings between the current research and Wu et al.'s (2010) work. The animations in this study had an accompanying narration – this was not present in Wu et al.'s (2010) materials. Kirschner et al.'s (2016) CTMA highlights how narrations can be a

source of ‘redundant’ information in assessment materials. The inclusion of this redundant information could have also been a source of extraneous assessment load (see Section 2.6.3). Participants with lower levels of prior knowledge find this type of cognitive assessment load particularly difficult to manage (Kirschner et al., 2016). In any case, the current study supports the literature’s assertions that the use of different types of visualisations may interact with other key variables to influence test-taker performance.

Item formats influence the psychometric characteristics of test items (Downing & Halydyna, 2006). Study 1A found evidence in favour of a context-dependent interaction between multimedia stimulus type and key item statistics. Five items encompassing a range of SR, CR and FR item types were identified as having notable differences in item discriminations across conditions (see Table 4.8). For four of these items, the use of dynamic multimedia objects in the item’s stimulus *improved* the item’s discriminatory ability. According to the CTMA (Kirschner et al., 2016), varying the level of coherence, contiguity or redundancy can allow an item to be more effective at discriminating between test-takers. Narrated animations allowed this to occur, thus offering a possible explanation for the higher discriminatory indices of these items in the dynamic condition. Some of the participants in Study 3 noted this stating that it was difficult to ‘manage’ the flow of information in some items with dynamic stimuli. This finding offers one reason why test developers should favour or at least consider using some dynamic items. This study found that dynamic stimuli improve discrimination, and improved discrimination is generally desirable according to Downing and Haladyna (2006). In contrast, the impact of multimedia type on item difficulties was more inconsistent and unpredictable. For example, one SR item in Study 1A had a higher item difficulty in the static condition than in the dynamic condition (Item P3). The reverse was true for the SR item in the next unit (Item G3). While these results cannot provide a clear and unambiguous answer as to the impact of multimedia stimuli on item functioning, it offers evidence that the use of static or dynamic objects in technology-based items has potential implications for key item statistics.

It should be noted that the discrepancies in item statistics between conditions were largely confined to items contained within the last two units of Study 1A’s instrument; ‘Blue Power Plant’ (‘P’), ‘Groundwater Extraction and Earthquakes’ (‘G’). Understanding why such discrepancies emerged requires a closer examination of the

stimulus content for these units. Carney and Levin (2002) argue that pictures may have at least five different functions in test items. 'Representational' visuals illustrate information from a verbal problem statement and include key details e.g. numbers, labels. Task critical information was included in the static stimuli for these two units through the inclusion of key words (e.g. 'Fault') and sequencing cues (e.g. numbered areas). While the other two units in Study 1A ('Meteoroid and Craters', 'Sustainable Fish Farming') also used representational features in their multimedia stimuli, more features were included in the 'Blue Power Plant' and 'Groundwater Extraction' units. When the pictures for these units were converted into animations for this study, even more task relevant information was available e.g. lights appearing *after* a turbine moved. This may have contributed, along with other factors like levels of prior knowledge, to these differences in item statistics between the two conditions. Eye movement data for these two units were examined to determine if any further insights to explain why items in these units functioned differently.

5.3.1.2 Impact on test-taker behaviour

Eye movement data revealed differences in test-taker behaviour between conditions. In particular, eye movement data showed that the *type* of multimedia stimulus used consistently affected test-takers' attentional behaviours with the interaction space of an item. For example, participants in the dynamic condition had significantly longer average fixation durations on an item's interaction space than those in the static condition. Taking into consideration the magnitude of the effect sizes calculated for the item-by-item comparisons conducted (see Table 4.11), participants spent more time fixating on information in the interaction space of an item when a *dynamic* stimulus was used. Participants in the dynamic condition, for the most part, also had an increased number of visits to an item's interaction space when compared with their counterparts in the static condition. Participants were more likely to leave and then return to the interaction space of an item when item stimuli were represented in a dynamic format. For five of the seven items investigated, the effect sizes for this difference in behaviour could be classified as moderate or large. Although the large standard deviations observed in these data may call into question the reliability of the means used, they do reflect the differences between participants on how long it took some to actually fixate on a relevant area when output was available to them. However, based on the evidence gathered in this

study, the modality of the item stimulus used did indeed affect test-taker attentional behaviour.

Taking into consideration the findings of Study 1A however, these attentional behaviours can also be used to support the assertion that narrated animations can act as a source of extraneous assessment load in TBAs. Participants in the dynamic condition of Study 1B visited an item's interaction space more often than those in the static condition. When these participants were visiting the item's interaction space, they had longer average fixation durations in this area than participants in the static condition. An increased number of visits to an AOI (i.e. the item's interaction space) is usually associated with increased interest and relevance (Carter & Luke, 2020). The act of returning to the area implies this. Yet, the act of having to leave the area is equally significant in an assessment context as it suggests that participants in the dynamic condition consulted another relevant or interesting area more often than those in the static condition. While the animation may have been more interesting to look at, increased attentional distribution (i.e. moving in and out of an item's interaction space) is usually associated with inefficient task performance, as shown in Brams et al.'s (2019) and Zagerman et al.'s (2016) work. Rayner et al.'s (2006) research on word identification in reading tasks also asserted that an increase in fixation durations indicates increased processing effort. Data from the c-RTAs also alludes to this, with many participants noting that the animations did not provide relevant information as efficiently as 'normal' stimuli e.g. text, images. Thus, it is suggested that the modality of the item stimulus acted as a source of extraneous load, and this had an effect on the attentional behaviour of test-takers, if not necessarily their ultimate performance (Kirschner et al., 2016).

5.3.2 Engaging with Simulation-Type Items

Study 2 investigated how test-takers engaged with simulation-type items derived from the PISA 2015 test of scientific literacy (OECD, 2018). The eye movement data collected showed that test-takers developed more efficient and effective test-taking strategies as their familiarity with these items increased. In Task 1, a longer time-to-first-fixation on relevant information for the Output phase⁶⁰ was associated with improved item performance but the inverse was true for all subsequent tasks and items. This

⁶⁰ The 'Output' phase began when the results of the first simulation were presented and continued until the participant had navigated away from the task

indicates that as familiarity with the simulation interface increased, so too did test-taker proficiency at identifying relevant information in the generated data which was then associated with improved item performance.

Eye movement data from Study 2 provided greater insight into the specific attentional behaviours that the simulation-type items elicited from test-takers. For example, when comparing the proportion of fixations on relevant and irrelevant information in the Output phase for each task, test-takers were shown to allocate more attention to the relevant areas of the simulation output than for the irrelevant⁶¹ areas. The effect sizes for these task-by-task comparisons were relatively large indicating considerable practical significance. This was the case for all tasks, with the exception of Task 1 (where test-takers were still familiarising themselves with these new item types). Task-by-task comparisons of eye movements in the Output phase by test-takers with varying levels of item performance⁶² also showed differing patterns of behaviour (see Table 4.26). Higher performing test-takers allocated much more attention to relevant information for Tasks 1-3. However, for Tasks 4 and 5, there was no difference in the proportion of fixations between higher and lower performing participants. According to the ECD framework (Hao & Mislevy, 2018), assessment tasks are presumed to elicit a particular response, behaviour or cognition that leads to an ‘observable’ outcome (e.g. selection of an option in an MCQ) from which inferences can be drawn. Information like this can be used to better articulate the relevance of key behaviours to the construct under investigation (Hao & Mislevy, 2018; Arieli-Attali et al. 2018). More specifically, by knowing the behaviours that preceded a certain action, test-developers can better describe how test-takers’ arrived at their final decision. This can then support the inferences that one can make. For example, understanding that the successful completion of a task is associated with test-takers paying more attention to relevant information (as in Tasks 1-3) would allow test-developers to indicate that the selection of a correct option in an MCQ occurred because test-takers were able to select or disregard relevant and irrelevant information as appropriate. This can inform inferences on the test-takers’ knowledge, skills or abilities.

⁶¹ As outlined in Chapter 3, ‘relevant’ and ‘irrelevant’ areas were categorised based on the content of the tasks’ items according to the PISA 2015 (OECD, 2016a) specifications.

⁶² Those test-takers who received Partial/No Credit and Full Credit for their responses.

Other log-file data variables, such as time-on-task and number of simulations run, were examined to better understand test-taker behaviour and performance with simulation-type items. There was no relationship between time-on-task and test-taker performance, except for Task 1. When time-on-task was investigated per phase, time spent by participants in the Orientation phase of Task 1 was negatively correlated with their performance on that task's item. That is, the *less* time spent in the Orientation phase of Task 1, the better test-taker performance was for the associated item. This is likely due to this being the least complex task in Study 2. If test-takers could grasp the requirements of this task quickly and execute their simulation correctly, they were likely able to perform well in the associated item. While some evidence suggests that the time test-takers spend on a task is positively correlated with item performance (e.g. Lee, 2018), the current study found no evidence in favour of this.

Studies have demonstrated that certain behaviours, such as the number of actions or simulations conducted by test-takers, play an important role in understanding task success and item performance (e.g., Greiff et al., 2015; Greiff et al., 2018). The PISA 2015 (OECD, 2016b) framework outlined the minimum number of simulations that test-takers should run to complete the task's associated items. The median number of simulations conducted by the participants for each task in Study 2 generally aligned with PISA 2015's minimum requirements. For the majority of these tasks, a moderate to strong, positive relationship between the number of simulations run and item performance was noted. For Task 4 however, participants ran more than the expected number of simulations. For this task, running more simulations than recommended was associated with improved item performance. Teig et al.'s (2020) work offers a compelling explanation as to why successful test-takers ran more simulations than expected for Task 4. In this task, test-takers had to run at least two simulations to select the highest temperature at which a person can run without getting heat stroke. While the most efficient approach to this task would be to start with the highest temperature setting and then vary it downwards in each simulation, stopping at the temperature where the runner is safe from heatstroke, Teig et al. (2020) argue that this efficiency does not always support sound scientific principles (e.g. null hypothesis testing, replicability). Indeed, in the case of this particular task, running slightly more than two simulations may have been indicative of thorough scientific reasoning. Generally speaking, a very high frequency of actions and simulations is likely to be associated with random behaviour, while the opposite may reflect poor task

engagement. However, the current study's findings, in conjunction with Teig et al's (2020) work, indicate that these assumptions should only be after certain contextual factors, such as subject area, have been considered.

The number of simulations generated by participants in Task 4 also offers some explanation as to why test-takers' attentional allocation behaviours appeared to differ on this task when compared with the other tasks in Study 2. For most tasks, high performers paid more attention to relevant areas than low performers. This difference in behaviour was not observed for Task 4. This indicates that while low performing test-takers who received partial or no credit for Task 4's items could identify the area where important information was housed, they could not select the pertinent information. This suggests that an increase in cognitive assessment load may have led to an increase in fixation counts. More successful participants however did not have as many fixations within the relevant areas. High performing participants appeared better able to process the large volume of data (i.e. germane cognitive assessment load; see Figure 2.10) associated with Task 4. A similar finding in relation to the proportion of fixations allocated by test-takers with differing categories of performance was also demonstrated with Task 5. In this task, test-takers were required to extrapolate beyond the data that could be directly collected through the simulation to complete the task's items. This increased level of task difficulty (i.e. intrinsic assessment load; see Figure 2.10) eliminated some of the differences in test-taker attention allocation behaviour that had been noted in Tasks 1, 2 and 3.

5.3.3 Interacting with Technology-Based Items

Three latent themes captured the nature of test-takers' interactions with technology-based test items. The first theme, 'Familiarisation', described how test-takers orientated themselves to the online testing environment and its test items. 'Sense-making', the second theme, described the different ways in which test-takers sought out and identified relevant information. The final theme, 'Making Decisions', highlighted the test-taking strategies that informed and affirmed test-takers' decisions. The data provided by participants for each of these themes allowed for a richer and more detailed understanding of the quantitative data collected. As discussed in the previous section, eye movement data gathered in Study 2 indicated that test-takers developed more efficient test-taking strategies for later, similar test items (e.g. identifying in advance where relevant information would be after a simulation). In the 'Familiarisation' theme, test-

takers acknowledged their increasing comfort and fluency with these items as the main reason for their change in attentional behaviour.

The 'Sense-making' theme provided similar insights. Sense-making is a concept originally derived from organisational theory (Weick, 1995) but has been applied in many different contexts to describe the actions preceding a judgement or decision e.g. social work (Platt & Turney, 2014). For this research, it is being used to describe the process through which test-takers interacted with all parts of a test item to gather the information needed to address the item's requirements. For example, interviewees revealed that they could 'make sense' of an item in the dynamic condition of Study 1B by gathering information from the item's stimulus through their auditory channel while simultaneously obtaining information on the contents of the interaction space through their visual channel. This aligns with Mayer's (2008; 2014) Cognitive Theory of Multimedia Learning. It also correlates with the quantitative data gathered in Study 1B (increased attention in the interaction space of an item) and highlights the role of the audio narration in supporting different test-taker behaviours. According to the qualitative data, other features of an animated item stimulus helped test-takers to 'make sense' of the test item e.g. moving turbines etc. However, the data also indicated that test-takers experienced different barriers in their search for relevant information when in the dynamic condition e.g. not being able to 'find' key words as easily. Test-takers also had preferences for different visualisations e.g. tables in Study 2, preference for static diagrams in Study 1B. These individual differences in the sense-making process further supports the idea that the design of items in TBAs can interact with a range of factors to affect test-taker behaviour.

Alemdag and Cagiltay (2018) assert that eye movement data can be interrogated to better understand an individual's decision making process. Unfortunately, this was not possible in the current study due to the spatial layout of the test items which were contrary to Carter and Luke's (2020) guidelines for such analyses. Instead, the data from the c-RTAs provided an insight into this process. Test-takers revealed that, wherever possible, they would 'transfer' the strategies that they would use in a standard pencil-and-paper exam to decide their final answer to a test item e.g. answering multi-part questions out of order etc. 'Post-decision' behaviours were easier to execute in an online environment for the test-takers. For example, 'double checking' their decisions rarely

resulted in an answer change but participants highlighted how easy it was to do in the online environment.

5.3.3.1 Interviewee 8

While focussing on the experiences of one participant can be problematic from a generalisability perspective, individual case studies can reveal important nuances in research findings (Coleman, 2019). This is particularly true when one considers the qualitative data provided by Interviewee 8 in Study 3. The data collected in Study 1A suggested that, on the whole, the use of dynamic or static stimuli may affect behaviour without systematically affecting performance. However, interview data indicated that test-takers may have a particular *preference* for certain types of stimuli. For example, Study 3 revealed that a number of test-takers had a marked preference for static stimuli. In the case of Interviewee 8, this preference for static stimuli was particularly noteworthy when analysed alongside the demographic and performance data collected from them. Interviewee 8 had a 'high' level of prior scientific knowledge based on their self-reported test scores. They completed the dynamic version of the TBA in Study 1B but only achieved a moderate score. Despite their high level of prior knowledge, it appears that Interviewee 8 could not manage the flow of information associated with the dynamic test items of Study 1B which they also acknowledged; *'I was just kind of overwhelmed with the video being there as well as the question'*.

Those with high levels of knowledge or skills on a topic should be able to manage extraneous cognitive assessment load according to Kirschner et al. (2016). It appears that the extraneous cognitive assessment load associated with dynamic items overwhelmed Interviewee 8 and that, perhaps, their reported level of scientific literacy was not accurately captured by previous assessments. However, in contrast to their performance in Study 1B, Interviewee 8 was the top performer in Study 2. Their performance in Study 2 suggests that they did indeed have a high level of scientific literacy. Consequently, the experiences of Interviewee 8 could be used as evidence in favour of the expertise reversal effect in testing, whereby techniques that support a novice (e.g. animations) can negatively impact an expert's performance (Kalygna & Renkl, 2010; Malone & Brünken, 2013). While this was not replicated with other participants who had similar levels of prior knowledge, its occurrence with a test-taker who performed exceptionally well on another instrument of scientific literacy (i.e. Study 2) does indicate that Kirschner et al.'s

(2016) concerns that poorly designed test items may mask a test-taker's expertise is warranted. Indeed, Interviewee 8's experiences suggests that there may be a certain type of test-taker who is disadvantaged by dynamic item types.

5.3.3.2 Test-Takers' views on TBAs in post-primary settings

Another theme, 'Recommendations', summarised participants' opinions on TBAs and their recommendations for the future deployment of the same in Irish post-primary schools, particularly for high stakes exams. The data gathered showed that test-takers were positively inclined towards online testing environments but with some important caveats, especially in relation to what subjects they were well suited for. Many of the participants identified mathematics as an unsuitable subject for online exams due to the need for 'rough work' when solving test items for this subject. This caution over the use of online tests for subjects that involve diagrams or formulae reflects the experiences of New Zealand post-primary students who responded negatively to the country's initial pilot of a TBA involving this subject (New Zealand Qualifications Authority; NZQA, 2014). Participants also had clear expectations on the usability of the testing platform with many of their recommendations unintentionally aligning with Molich and Nielson's (1990) widely cited design principles. Other recommendations on what tools and skills test-takers would be needed in advance of a widespread introduction of TBAs for high-stakes exams, including touch typing skills, were also provided and will be discussed in greater detail in Section 5.7.1.

5.4 Conclusions

Taken together, the findings from Study 1A and Study 1B show that the type of multimedia stimulus used in an item can affect test-taker attentional behaviour *without* necessarily impacting overall performance. While differences in eye movement data between conditions were observed, no associated differences in test scores were found. However, dynamic stimuli with audio narrations appeared to precipitate eye movements associated with extraneous cognitive assessment load. The data gathered in Study 3 supports this conclusion, with test-takers themselves admitting that the dynamic stimuli sometimes made it difficult to manage the 'flow' of information. It should also be noted that differences in multimedia stimuli also led to differences in key item statistics, including item discriminations, for certain test items. Such findings support claims that

different item stimuli may advantage or disadvantage certain groups of test-takers (e.g. Kirschner et al., 2016; Moon et al., 2019). These conclusions, along with similar findings by Moon et al. (2019) and Arslan et al. (2019), provide a basis for a much-needed discussion on how different test item formats may affect test-taker behaviour and item functioning.

The data gathered in Study 2 provides further insight into *how* test-takers engage with simulation-type items. Based on the data gathered, increasing test-taker familiarity with simulation-type items can affect test-taker attentional behaviour leading to more effective test-taking strategies. More successful test-takers directed significantly more of their attention to the relevant areas of the simulation output. However, generating large volumes of data (Task 4) or having to extrapolate from 'missing data' (Task 5) appears to disrupt the predictive properties of these behaviours. Examination of other process data variables (e.g. time-on-task, number of simulations run) showed that some of the most common interpretations ascribed to frequencies of these behaviours (e.g. Grief et al., 2015) are context and subject specific.

The qualitative data gathered for this research identified some of the key interactions post-primary test-takers exhibited when engaging with technology-based items. In particular, the interview data highlighted how different features of test items, like multimedia stimulus and item layout, can support or undermine the sense-making process. Furthermore, the use of an online testing environment facilitated certain test-taking strategies e.g. reviewing responses. The data gathered also allows for more specific conclusions about the ideas and preferences of post-primary students in relation to the introduction of TBAs in Irish post-primary schools to be drawn. While test-takers are positively disposed to the introduction of online tests, they have a number of key recommendations to support such a significant change in the Irish education system.

The conclusions from the study are presented visually in Figure 5.1, numbered from one to eight.



Figure 5.1 Conclusions

5.5 Study Strengths and Limitations

The assessment materials used for this research were derived from the publicly available test items on scientific literacy that were designed for PISA 2015. These items are based on a clear framework of scientific literacy (see Appendix B) and underwent a rigorous development process (OECD, 2016b). They are considered to be a good measure of post-primary students' scientific literacy (OECD, 2017; 2016c). Their use in this research with an ecologically valid sample within the age range of the students that participated in PISA 2015 increases confidence in this study's findings. Much of the other research in the field (e.g. Wu et al., 2010) is based on researcher designed tasks and tests which has many inherent limitations (Coleman, 2019). Furthermore, while the use of self-reported grades and scores can often be a cause for concern due to the risks of social desirability or inaccurate recall (Coleman, 2019), this does not appear to be the case for this research. Participant performance in Study 1A, Study 1B and Study 2 was in line with their reported levels of prior scientific knowledge, offering strong criterion validity evidence for the materials used in this study (AERA et al., 2014; Fraenkel & Wallen, 2006). Convergent validity evidence was also demonstrated by the strong, positive association calculated between test-taker performance on Study 1B and Study 2.

Regardless of the strengths of the current study, it is important to acknowledge that everything should be considered within the context of the limitations of the research. The first limitation of the study results from the way in which the sample was selected. Schools' participation in Study 1A was entirely voluntary. The schools and students that did engage with the research project often had a particular interest in digital assessment⁶³ thus making the risk of a self-selection bias likely (Coleman, 2019). Those who volunteered their schools' involvement in the study, as well as the participants themselves, may have had different characteristics from those who did not participate. This may have skewed the sample and the results. However, due to ethical constraints and other contextual reasons, this was a limitation that could not be avoided.

Data collection occurred between October and December 2020 in Irish post-primary schools. At this point, the COVID-19 pandemic was ongoing in Ireland and a number of restrictions regarding non-teaching visitors to school settings were in place

⁶³ The majority of the schools involved in this research were piloting the Leaving Certificate Computer Science curriculum in their schools. A digital exam for this subject, the first of its kind in Ireland, was planned for May 2021 at the time of data collection.

(DES, 2020). As a result, direct access to the schools and participants involved in Study 1A was not possible. Consequently, participants completed the TBA under the direction of their class teacher rather than the researcher. While standardised instructions were used (Appendix I), the standardised administration of the testing materials for Study 1A cannot be fully guaranteed. Difficulties in accessing participants also impacted the final sample sizes for this research. The sample sizes for Study 1B and Study 2 were not large enough to obtain statistical power at the recommended .80 level (Cohen, 1988). This issue was particularly pronounced when conducting sub-group analyses e.g. analysis of test-taker performance according to condition and prior levels of scientific literacy. Differential Item Functioning (DIF) analysis would have been able to determine whether items in different conditions advantaged or discriminated against certain test-takers but the sample sizes available were too small, even taking into consideration advances in the use of DIF with small sample sizes e.g. Belzak (2019). The small sample size of the studies conducted for this research may call into question the robustness of the conclusions drawn.

A common set of test items were used to ensure construct invariance for the two experimental conditions in Study 1A and Study 1B. The animated videos used in the dynamic condition were based on the written text contained in the static condition. While the content of the test items was held constant, it can only be assumed that the same constructs were being measured across the two conditions. This is unsatisfactory, given that Lievens and Sackett's work (2006) demonstrated that different multimedia stimuli can affect construct representation in SJTs involving interpersonal situations. Given the sample sizes involved, statistical evidence using the multiple group approach to confirmatory factor analysis (e.g. Chan & Schmitt, 1997) could not be obtained to demonstrate construct invariance across the two test formats in this research. It is in this study, whether or not the use of animations can affect the construct being investigated by an item. This is a significant limitation but it does provide a clear avenue for future research (see Section 5.7.2). For example, Study 1A showed that high performing students had improved scores on items with dynamic stimuli. Based on this finding, a second construct may have been unintentionally assessed when items used dynamic stimuli e.g. participants' ability to manage and process information. To determine if this is the case, further research that directly investigates construct representation in items of different designs is required.

For Study 2, simulation-type items were used. These are new and innovative item types and research on them in assessment contexts is only beginning to emerge e.g. Teig et al. (2020). While previous literature has identified some of the different phases associated with completing these tasks e.g. Teig et al. (2020), Greiff et al. (2016; 2018), Pedaste et al. (2015), there is little consistency regarding their terminology and operational definitions. Indeed, none of the terms that are currently being debated in the field were deemed appropriate to Study 2. The ‘Orientation’ and ‘Output’ phases used to categorise test-takers’ interactions in Study 2 do not align to any particular framework on complex problem solving or inquiry based assessments. Instead, the labels used in Study 2 refer to the general ‘actions’ undertaken by test-takers when engaging with TBAs (Lindner et al., 2017a). If these actions had been mapped onto a conceptual framework at the time of the study’s inception, the generalisability and replicability of Study 2 would be far greater.

A large volume of process data was generated for Study 1B and Study 2. In line with Carter and Luke’s (2020) recommendations, a ‘preregistration’ approach was applied, whereby specific hypotheses and analysis plans were outlined prior to data collection and analysis. While this safeguarded against some of the more common issues with process data, such as data fishing and HARKing⁶⁴, other analyses that may have revealed relevant information were not conducted. For example, a growth curve analysis would have revealed if and how attentional differences in Study 1B emerged using time as a predictor (Dink & Ferguson, 2015). An onset contingent analysis would have provided a more in-depth examination to test-takers’ attentional reactions in Study 2 (Dink & Ferguson, 2015). Time-based scanpath sequencing would have also provided more detail on the sequencing of test-takers’ engagement with simulation-type items (Zhegallo & Marmalyuk, 2015). The use of a preregistration approach precluded the use of such analyses. While unfortunate, it must be acknowledged that these limitations can be used as a catalyst for future explorations on the practical implications of the research presented here.

⁶⁴ Hypothesizing after results are known (Field, 2018).

5.6 Theoretical and Practical Implications

This research has responded to calls for a more in-depth examination of item designs for TBAs (Bryant, 2017; Kirschner et al., 2016; Malone & Brünken, 2013), while also adding to the growing literature on simulation-type items (Teig et al., 2020). Furthermore, rather than solely relying on the accuracy of test-taker responses, the process and qualitative data collected in this study provided additional insights as to *how* the responses were produced. Compared to previous research e.g. Teig et al. (2020), Lindner et al. (2017a; 2017b; 2020), the triangulation of data in this study allowed for a richer and more complete view on the use of different items in TBAs. This has several theoretical and practical implications for the field.

The current research is one of the first to explore the use of different types of multimedia stimuli in assessment contexts. Previous research, such as that conducted by Lindner et al. (2020), generally compared test-taker performance on text-only and text-image test items. This research explored the use of animations in test items, something that has been less commonly explored in the field. This research will supplement what test-developers already know about the use of animations in assessments and should also provide some much-needed foresight on what should be considered when including such dynamic stimuli. For example, this research highlighted what other factors multimedia stimuli could interact with to modulate item discriminations, such as the use of audio narrations, test-takers' levels of prior knowledge and the use of representational features in the multimedia stimulus. Such information can be used to encourage a more robust approach to the design of item stimuli.

The range of data gathered in this research also offers an explanation as to *why* multimedia stimuli can vary item discriminations. The eye movement data collected for Study 1B indicated that dynamic stimuli often caused attentional behaviours that are commonly associated with extraneous processing (e.g. Zagerman et al., 2016). Understanding that stimulus modality can be a source of extraneous assessment load can offer test-developers a way to create items with an 'optimal' amount of extraneous cognitive assessment load, a key aim when developing items for TBAs (Kirschner et al., 2016). Knowing how to create items with this ideal level of cognitive assessment load may allow for items with greater discriminatory ability, a desirable feature of test items (AERA et al., 2014). While further research is required to better understand how this can be achieved and what factors need to be considered (e.g. audio narration), these findings

have highlighted some item design features that can be exploited to support item quality. The findings of Study 1A and 1B will support the development of a more robust cognitive theory of multimedia assessment, a much needed requirement in the field (Kirschner et al., 2016).

Study 2 adds a new contribution to the growing literature on the use of simulation-type items in educational assessments. In particular, it showed how increased familiarity with these item types can support the development of more effective test-taking strategies e.g. identifying relevant information. This finding has two significant implications. Firstly, it suggests that new and innovative item types, such as simulations, in TBAs may indeed introduce construct-irrelevant variance into the testing process. Test-taker unfamiliarity and uncertainty with the item itself, rather than the construct being investigated, affected test-taker behaviour. This difference in behaviour could impact how scores are interpreted. Therefore, this study justifies previous concerns on the introduction of construct-irrelevant variance to testing contexts as a result of 'innovative' items as highlighted by Moon et al. (2019), Bryant (2017) and Russell (2016). However, the study also shows that the use of authentic 'practice' items could address this issue. While test-takers in Study 2 did have an opportunity to become familiar with the simulation controls before beginning the unit, this was not considered a real 'practice' item by test-takers. This is likely because test-taker engagement with the controls involved test-takers 'exploring' the simulation controls as they saw fit rather than actually trying to complete a task. Although Kirschner et al. (2016) argue that 'pre-training' should be avoided in assessment contexts (see Table 2.4), this research suggests that practice items should be provided. However, practice items need to balance being authentic without providing too many cues to test-takers. Understanding what constitutes 'useful' practice for simulation-type items may support the design of better TBAs.

Data collected in Study 2 provided an in-depth description of the behaviours of test-takers when completing simulation-type items. Process data in the form of eye movement, time-on-task and number of simulations run helped to create these descriptions. As a result, evidence on the potential relationships between certain behaviours and performance was gleaned e.g. time-to-first-fixation and item success. Evidence advocating a context and item specific interpretation of these behaviours was also found e.g. Task 4. The ECD e-assessment framework notes that it is important to

explicitly link relevant competencies, skills or knowledge with behaviours (Hao & Mislevy, 2018). The information gathered in this study demonstrates how this can be achieved using process data. It also underscores that validation studies involving process data for TBAs ensure that appropriate inferences are made (Embretson, 2016; Lane, 2017).

Study 3 also progresses the literature in two ways. To begin, the use of c-RTAs using eye movement data as cues has, traditionally, been limited to usability studies (Elbabour et al., 2017; Olsen et al., 2010). The current study has demonstrated how c-RTAs can be deployed in non-usability studies to provide useful data for researchers across a range of disciplines. Of greater consequence are the insights provided by the post-primary students on TBAs. TBAs that are designed and used for this cohort are often based on research conducted with university students (Bryant, 2017) which may be problematic for generalisability reasons, especially given that literature has previously highlighted how test-taker interactions with TBAs can vary significantly by age e.g. Wu et al. (2015). By obtaining the views and perspectives of these test-takers, the current research is gathering data that is directly relevant to this cohort. For example, test-taker disclosures highlighted how the deployment of dynamic stimuli for certain items assisted in the all-important 'sense-making' process. Other insights in relation to TBA designs e.g. layout of items, use of white space were also revealed. The propensity with which test-takers' engaged with 'reviewing' behaviours after completing an item in a TBA also appears to be a relatively new finding in the literature. These insights add some much needed depth and breadth to the field's understanding of test-taker interactions with TBAs. The views of these test-takers are also a resource for policymakers and researchers who can use this study as a primer for introducing TBAs to a curriculum.

5.7 Recommendations

5.7.1 Policy and Practice

This research demonstrated how item formats can affect test-taker behaviour and test performance as well as the psychometric properties of test items. It has shown that variations in item design (e.g. use of dynamic stimuli) can result in different test-taker behaviours and experiences. This item specific information should be utilised by those involved in the design of TBAs to ensure a more valid assessment design in line with the recommendations contained in the *Standards* (AERA et al., 2014) and the ECD framework

(Hao & Mislevy, 2018). When accompanied by further research, such as that suggested below, this may eventually allow for the development of more specific item writing guidelines for TBAs while also helping to further develop and expand Kirschner et al.'s (2016) CTMA.

Research as to how post-primary education systems can effectively deploy TBAs for high stakes exams is extremely active (e.g. Lehane, 2019). The current study offers some practical advice to assist in this endeavour. Based on the experiences of other countries like New Zealand, Lehane (2019) strongly advocated for the use of familiarisation activities to support the orientation of Irish post-primary students to TBAs. The author specifically cites the importance of having all possible item types included in these activities. Based on the eye movement and qualitative data collected, this research also supports the use of familiarisation activities for TBAs involving post-primary students. As seen in this research, they are necessary to support effective test-taking strategies. By having some familiarity with the content and layout of the testing environment in advance of using it in a high stakes context, test-taker ease with the TBA is likely to be increased. Construct irrelevant variance caused by test-taker anxiety or uncertainty is likely to be reduced (Wise, 2018). Furthermore, the amount of time it took participants to complete the dynamic version of the test in Study 1A was statistically and practically longer than those in the static condition. This is not surprising given that the animations used in the dynamic condition ranged in duration from 22 seconds to 60 seconds. Multiple viewings of these animations would also contribute to the longer testing time associated with the dynamic condition. Therefore, the use of differing multimedia stimuli in a TBA may require different time limits than previous iterations of TBAs or paper-based tests.

A number of specific recommendations regarding design of TBAs for use with Irish post-primary students were identified in the qualitative data. Test-takers in this study believed that a sound interface design was essential for success in high stakes exams and assessments. This was also something that Lehane (2019) also recommended, citing Molich and Nielson's (1990) seminal work as a starting point for such efforts. Similarly, Harms and Adams (2008) asserted that each component of an online interface, regardless of the prospective device, industry or purpose, must be designed 'with consideration of the knowledge, expectations, information requirements, and cognitive capabilities of all possible end users' (p. 4). Therefore, the interface design of a TBA should take into

consideration the specific needs of students in an online testing environment. The test-takers in this research highlighted some of these needs, including ‘warnings’ if a question had been forgotten, the freedom to navigate between items and the ability to review their final responses. By combining the recommendations of these test-takers with the widely endorsed usability heuristics for interface design (Molich & Nielson, 1990), the design of high stakes TBAs for post-primary students will be much improved. Furthermore, given the insights obtained from the participants involved in this research, the co-production of high stakes TBAs *with* post-primary students would also be particularly beneficial. Research that examined the co-production of health interventions in English post-primary schools (Ponsford et al., 2021), demonstrated that post-primary students are well placed to highlight facilitators or barriers to implementation and acceptability. They can also identify potential unintended consequences and ways of addressing these. This would be very valuable to the field of educational TBAs.

Russell’s (2016) TEI utility framework is designed to ‘weigh the costs and benefits of employing a given response interaction methodology to measure the knowledge, skill, or ability of interest’ (p. 24). Engaging in a cost-benefit analysis is also relevant for an item’s stimulus. In relation to the use of dynamic stimuli, it should be determined whether animated item stimuli can justify the resources for their development. Based on this study, there does appear to be some justification for the cost and efforts associated with their use e.g. improved discrimination, source of cognitive assessment load. However, to engage in an accurate cost-benefit analysis, further research is required to fully understand when it is most beneficial to deploy animated stimuli in test-items.

5.7.2 Future Research

The limitations of the current research can help guide the path of future research. In particular, future research should ensure that participant sample sizes are sufficiently large to detect even small effect sizes. This would allow for a wider range of analyses e.g. DIF, which would, ideally, identify if and when different item design decisions affect different categories of test-takers. Work by Belzak et al. (2019) could be applied in the identification of minimum sample sizes for such work. Regardless of these small sample sizes, the current research demonstrated the potential contribution of eye movement data as a source of process data for educational assessments. Future research should use eye movement data to further expand the field’s knowledge of test-takers’ attentional

behaviours and cognitive processes when completing technology-based items. For example, this research used a relatively simple analysis of summary AOI statistics to test specific hypotheses on item processing. Future research should pay greater attention to the *sequence* of test-takers' eye movements. As noted by Oranje et al. (2017), examining sequence information could reveal certain patterns that may provide meaningful explanations of test-takers' strategies. The cost of eye-tracking systems and the development of psychometric models for analysis are becoming less prohibitive (Alemdag & Cagiltay, 2018) so this is a viable avenue for future research.

Research indicates that some of the item types examined in this research (e.g. SR, FR, CR) offer more construct fidelity than others (e.g. Russell & Moncaleano, 2019). While this research did not conduct direct comparisons between different categories and versions of items in terms of construct representation, the need for research that does so is still required. As seen in Study 3, test-takers in the dynamic condition utilised different skills than those in the static condition when completing test items e.g. interpreting dynamic representations of energy conversion versus inferring energy conversions from a static image. Furthermore, the use of an audio narration in the animations resulted in different test-taker attentional behaviours. Understanding test-taker cognition, behaviour and performance in alternative, construct equivalent item formats could further inform test score interpretations and the psychometric properties of items. As discussed by Grover et al. (2017), research like this may lead to better 'scoring rules' and more appropriate inferences between test-taker behaviours and competencies. Findings from this research study also suggest that examining the relationship between different item types and formats with test-taker background variables would also be a worthwhile area for future research. As noted by Russell (2016), conducting such research would ensure that the 'effort' associated with developing different certain technology-based items would be appropriate and beneficial from a validity perspective. This could ultimately lead to the development of evidence based item deployment and design decisions, an important goal in the field (Bryant, 2017).

Based on the qualitative data collected, it also appears that further research regarding the readiness of post-primary test-takers for digital assessments is needed, particularly whether or not they have the digital literacy skills necessary to effectively engage with TBAs. Fraillon et al. (2013) define digital literacy as the ability to use digital resources as a receptive and productive tool to collect, create, transform, and safely use

information. While post-primary students are often erroneously considered to be ‘digital natives’ (Prensky, 2001), research has found that despite early and prolonged exposure to technology, they often lack the skills necessary for effective and critical technology use (e.g. Lazonder et al., 2020). This aligns with the data gathered in Study 3 whereby participants acknowledged their difficulties in generalising behaviours across digital environments (e.g. dragging-and-dropping objects) and self-identified limitations in their digital literacy skills (e.g. typing proficiency). The findings from this study reveal a new and fertile ground for future digital literacy research, particularly its relationship to test-takers’ confidence and competence with digital assessment techniques. For example, future research should explore if and how the designated training of digital literacy skills could support comfort and performance in TBAs.

5.8 Epilogue

While assessment can take many forms, end-of-course high stakes exams are often the dominant form of assessment in many second-level systems (Keane & McNerney, 2017). In recent times, high-stakes computer-based exams and TBAs have become more commonplace in this sector e.g. Partnership for Assessment of Readiness for College and Careers (PARCC; United States), National Certificate of Educational Achievement (NCEA; New Zealand). While TBAs like these are transforming the assessment landscape, it is critical that those involved in assessment design avoid adopting a “techno-centric” mindset, whereby technology is used simply because it is available. The aim should be to harness technology in such a way that truly enhances the assessment process, by expanding the possibilities of what can be assessed, or the extent of the inferences that can be made from the assessment.

Given the context in which the current study took place, and taking into consideration the recent initiatives involving TBAs for the Leaving Certificate Examination (SEC, 2021) as well as the ‘Digital Strategy for Schools’ (DES, 2021), the findings of this research will be particularly pertinent to Irish educational policy makers. However, they also have relevance well beyond the Irish context. In particular, this research provides test-developers worldwide with insights as to how item features and test-taker attentional behaviours influence the psychometric properties of assessments and the inferences drawn from the data they provide. It has been argued elsewhere that students can, eventually, ‘escape’ the effects of poor teaching (Boud, 1995). However, it

is more difficult to escape the effects of poor assessment, particularly if those assessments are used as gatekeepers to further education or employment. Therefore, it is incumbent on everyone concerned to ensure that TBAs are the best they can be. Future generations of test-takers should expect nothing less.

Reference List

- American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), & Joint Committee on Standards for Educational and Psychological Testing (JCSEPT). (2014). *Standards for Educational and Psychological Testing*. AERA.
- Alemdag, E., & Cagiltay, K. (2018). A systematic review of eye tracking research on multimedia learning. *Computers & Education*, 125, 413–428.
<https://doi.org/10.1016/j.compedu.2018.06.023>
- Almond, R.G., Jeon Kim, Y., Velasquez, G., & Shute, V.J. (2014). How task features impact evidence from assessments embedded in simulations and games. *Measurement: Interdisciplinary Research and Perspectives*, 12(1), 1-33.
<https://doi.org/10.1080/15366367.2014.910060>
- American Psychological Association (APA). (2019). *Criterion validity* [dictionary].
<https://dictionary.apa.org/criterion-validity>.
- Anderson, L.W. (Ed.), Krathwohl, D.R. (Ed.), Airasian, P.W., Cruikshank, K.A., Mayer, R.E., Pintrich, P.R., Raths, J., & Wittrock, M.C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (complete edition). Longman.
- Arieli-Attali, M., Ward, S., Thomas, J., Deonovic, B., & von Davier, A. (2019). The expanded Evidence-Centered Design (e-ECD) for learning and assessment systems: A framework for incorporating learning goals and processes within assessment design. *Frontiers in Psychology*, 10, 853. <https://doi.org/10.3389/fpsyg.2019.00853>
- Arslan, B., Jiang, Y., Gong, T. & Keehner, M. (April, 2019). *The effect of drag-and-drop item type design on test-takers' performance and strategy use* [paper presentation]. Annual Conference of American Educational Research Association (AERA), Toronto.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- Baddeley, A.D. (1992). Working memory. *Science*, 255(5044), 556–559. doi: 10.1126/science.1736359
- Baker, R., & Clarke-Midura, J. (June, 2013). *Predicting successful inquiry learning in a virtual performance assessment for science* [paper presentation]. 21st International Conference on User Modeling, Adaptation, and Personalization, Rome.
- Bakia, M., Murphy, R. Anderson, K. & Trinidad, G. (2011). *International experiences with technology in education: Final report*. U.S. Department of Education, Office of Educational Technology and the Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service.
<https://oerknowledgecloud.org/sites/oerknowledgecloud.org/files/iete-full-report.pdf>

- Belzak, W.C.M. (2019). Testing differential item functioning in small samples. *Multivariate Behavioural Research*, 55(5), 722-747.
<https://doi.org/10.1080/00273171.2019.1671162>
- Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education*, 39, 370-407. doi:10.3102/0091732X14554179
- Bloom, B.S. (Ed.), Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. David McKay.
- Boud, D. (1995). Assessment and learning: Contradictory or complimentary? In Knight, P. (Ed), *Assessment for Learning in Higher Education* (pp. 35-48). Kogan Page.
- Brams, S., Ziv, G., Levin, O., Spitz, J., Wagemans, J., Williams, A. M., & Helsen, W. F. (2019). The relationship between gaze behavior, expertise, and performance: A systematic review. *Psychological Bulletin*, 145(10), 980-1027.
<https://doi.org/10.1037/bul0000207>
- Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101, DOI: 10.1191/1478088706qp063oa
- Bridgeman, B., Lennon, M.L. & Jackenthal, A. (2003). Effects of screen size, screen resolution, and display rate on computer-based test performance. *Applied Measurement in Education*, 16, 191-205. doi:10.1207/S15324818AME1603_2
- Bruckmaier, G., Binder, K., Krauss, S., & Kufner, H-M. (2019). An eye-tracking study of statistical reasoning with tree diagrams and 2×2 tables. *Frontiers in Psychology*, 10, 632, 1-28. <https://doi.org/10.3389/fpsyg.2019.00632>
- Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardized tests. *Practical Assessment, Research & Evaluation*, 22(1), 1-5.
<https://pareonline.net/getvn.asp?v=22&n=1>
- BTL (2018). *Surpass* [website]. <http://www.btl.com/surpass>.
- Butcher, K. (2014). The multimedia principle. In R. Mayer (Ed.) *The Cambridge Handbook of Multimedia Learning* (2nd Edition) (pp. 174-205). Cambridge University Press.
- Carney, R.N., & Levin, J.R. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review*, 14, 5-26.
<https://doi.org/10.1023/A:1013176309260>
- Carter, B. & Luke, S. (2020). Best practices in eye tracking research. *International Journal of psychophysiology*, 155, 49-62. <https://doi.org/10.1016/j.ijpsycho.2020.05.010>
- Chadwick, R., McLoughlin, E., & Finlayson, O. (2016). *Assessment and development of scientific literacy at second level*. National Council for Curriculum and Assessment. https://www.ncca.ie/media/1477/resanddev_rchadwick.pdf.

- Chan, D., & Schmitt, N. (1997). Video-Based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, 82, 143-159. <https://dx.doi.org/10.1037/0021-9010.82.1.143>
- Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4), 293-332. https://doi.org/10.1207/s1532690xc0804_2
- Clark, R., & Mayer, R. (2016). *e-Learning and the science of instruction* (4th Edition). Wiley.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Edition). Lawrence Erlbaum.
- Coleman, R. (2019). *Designing experiments for the social sciences: How to plan, create and execute research using experiments*. SAGE Publications.
- Creswell, J. (2014). *Research design: Qualitative, quantitative and mixed methods approaches* (4th Edition). SAGE Publications.
- Creswell, J. & Plano-Clark, V. (2011). *Designing and conducting mixed methods research*. SAGE Publications.
- Cronbach, L.J. (1988). Test validation. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd Edition; pp. 443-507). American Council on Education.
- Dalrymple, K., Manner, M., Harmelink, K., Teska, E., & Elison J. (2018). An examination of recording accuracy and precision from eye tracking data from toddlerhood to adulthood. *Frontiers in Psychology*, 9, 803. <https://doi.org/10.3389/fpsyg.2018.00803>
- Dancy, M. H., & Beichner, R. (2006). Impact of animation on assessment of conceptual understanding in physics. *Physical Review Special Topics - Physics Education Research*, 2(1), 1-7. <https://doi.org/10.1103/PhysRevSTPER.2.010104>
- Davidsson, E., Allerup, P., Davidsson, E., Sørensen, H., & Allerup, P. (2012). Assessing scientific literacy through computer-based tests: Consequences related to content and gender. *NorDiNa*, 8(3), 269-282. <https://doi.org/10.5617/nordina.533>
- de Jong, T. (2010). Cognitive load theory, educational research, and instructional design: Some food for thought. *Instructional Science*, 38(2), 105-134. <https://doi.org/10.1007/s11251-009-9110-0>
- DeLeeuw, K., & Mayer, R. (2008). A comparison of three measures of cognitive load:

- Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology*, 100(1), 223-234. <http://dx.doi.org/10.1037/0022-0663.100.1.223>
- Department of Children and Youth Affairs. (2015). *Children First Guidelines*. Stationary Office. https://www.dcy.gov.ie/documents/Publications/Ethics_Guidance.pdf
- Department of Children and Youth Affairs. (2012). *Guidance for developing ethical research projects involving children*. Stationary Office. https://www.dcy.gov.ie/documents/Publications/Ethics_Guidance.pdf
- Department of Education and Skills (DES). (2018). *Post-primary database of schools*. <https://www.education.ie/en/Publications/Statistics/Data-on-Individual-Schools/post-primary/post-primary-schools-2019-2020.xlsx>
- Department of Education and Skills (DES). (2020). *COVID-19 response plan for the safe and sustainable operation of post-primary schools*. <https://assets.gov.ie/83312/6c36aaac-22fc-44fd-a4be-88cea4db82d6.pdf>
- Department of Education and Skills (DES). (2021). *Digital Strategy for School Consultation Framework* [website]. <https://www.education.ie/en/Schools-Colleges/Information/Information-Communications-Technology-ICT-in-Schools/digital-strategy-for-schools-consultation-framework.html>
- Dink, J. W., & Ferguson, B. (2015). *eyetrackingR: An R library for eye-tracking data analysis* [Computer software]. <http://www.eyetrackingr.com>.
- Dolan, R. P., Goodman, J., Strain-Seymour, E., Adams, J., & Sethuraman, S. (2011). *Cognitive lab evaluation of innovative items in mathematics and English language arts: Assessment of elementary, middle, and high school students research report*. Pearson Education. <http://www.pearsonassessments.com/research>.
- Downing, S. & Haladyna, T. (2006). *Handbook of Test Development*. Lawrence Erlbaum Associates.
- Dublin City University (DCU). (2019). *Research ethics* [webpage]. <https://www.dcu.ie/researchsupport/researchethics.shtml>.
- Educational Research Centre (ERC). (2018, August 31). *New primary tests will be available soon!* <http://trythetests.erc.ie/article/NewPrimaryTests>
- Eisenhart, M. (October, 1991). *Conceptual frameworks for research circa 1991: Ideas from*

- a cultural anthropologist; Implications for mathematics education researchers* [paper presentation]. 13th Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education (pp. 202-219).
- Elbabour, F., Alhadreti, O., Mayhew, P. (2017). Eye tracking in retrospective think-aloud usability testing: Is there added value? *Journal of Usability Studies*, 12(3), 95-110. http://uxpajournal.org/wp-content/uploads/sites/8/pdf/JUS_Elbabour_May2017.pdf
- Embretson, S. (2016). Understanding examinees' responses to items: Implications for measurement. *Educational Measurement: Issues and Practice*, 35(3), 6-22. <https://doi.org/10.1111/emip.12117>
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th Edition). SAGE Edge.
- Fraenkel, J. & Wallen, N. (2006). *How to design and evaluate research in education* (6th ed.). McGraw-Hill.
- Fraillon, J., Schulz, W., & Ainley, J. (2013). International computer and information literacy study: Assessment framework. International Association for the Evaluation of Educational Achievement (IEA). https://doi.org/10.1007/978-3-030-19389-8_2
- Fulcher, G. & Davidson, F. (2008). *Language testing and assessment: An advanced resource book*. Routledge.
- Ramos-Gameiro, R., Kaspar, K., König, S.U., Nodholt, S. & König, P. (2017). Exploration and exploitation in natural viewing. *Scientific Reports*, 7, 2311. <https://doi.org/10.1038/s41598-017-02526-1>
- Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education*, 91, 92– 105. <https://doi.org/10.1016/j.compedu.2015.10.018>
- Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior*, 61, 36– 46. <https://doi.org/10.1016/j.chb.2016.02.095>
- Greiff, S., Molnár, G., Martin, R., Zimmermann, J., & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments: A latent class approach. *Computers & Education*, 126, 248– 263.

<https://doi.org/10.1016/j.compedu.2018.07.013>

- Gregory, S. (2019, August 1). *Why responsive design is important and google approved*. FreshSparks. <https://freshsparks.com/why-responsive-design-is-important/>
- Groff, J. (2018). The potentials of game-based environments for integrated, immersive learning data. *European Journal of Education*, 53(2), 188–201. <https://doi.org/10.1111/ejed.12270>
- Grover, S., Bienkowski, M., Basu, S., Eagle, M., Diana, N., & Stamper, J. (2017). A framework for hypothesis-driven approaches to support data-driven learning analytics in measuring computational thinking in block-based programming. *ACM Transactions on Computing Education*, 17(3). <https://doi.org/10.1145/3105910>
- Haladyna, T.M. (1999). *Developing and validating multiple-choice test items*. Erlbaum.
- Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and validating test items*. Taylor & Francis.
- Hao, J. & Mislevy, R. (2018). The evidence trace file: A data structure for virtual performance assessments informed by data analytics and evidence-centered design. *ETS Research Report Series*, 2018, 1-16. doi:10.1002/ets2.12215
- Harms, M. & Adams, J. (2008). *Usability and design considerations for computer-based learning and assessment* [paper presentation]. American Educational Research Associations (AERA), New York. <https://pdfs.semanticscholar.org/3661/cbe8b6d8fec7ad37648164d02f2a80d47960.pdf>
- Hessels, R.S., Kemner, C., van den Boomen, C. & Hooge, I. (2016). The area-of-interest problem in eyetracking research: A noise-robust solution for face and sparse stimuli. *Behavioural Research*, 48, 1694–1712. <https://doi.org/10.3758/s13428-015-0676-y>
- Ho, V., Harris, P., Kumar, R., & Velan, G. (2018). Knowledge maps: A tool for online assessment with automated feedback. *Medical Education Online*, 23(1). <https://doi.org/10.1080/10872981.2018.1457394>
- Holmqvist K., Nyström M., Andersson R., Dewhurst R., Jarodzka H., & Van de Weijer J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.
- Howitt, D. & Cramer, D. (2005). *Introduction to Research Methods in Psychology*. Pearson Education.

- Hubley, A., & Zumbo, B. (2017). Response processes in the context of validity: Setting the stage. In B.D. Zumbo & A.M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 1–12). Springer International Publishing.
- Hyönä, J. (2010). The use of eye movements in the study of multimedia learning. *Learning and Instruction*, 20(2), 172-176.
<http://dx.doi.org/10.1016/j.learninstruc.2009.02.013>
- Iseli, M., Koenig, A., Lee, J. & Wainess, R. (2010). *Automated assessment of complex task performance in games and simulations* (CRESST Research Rep. No. 775). CRESST.
<https://files.eric.ed.gov/fulltext/ED512656.pdf>
- Jamet, E., Gavota, M., & Quaireau, C. (2008). Attention guiding in multimedia learning. *Learning and Instruction*, 18(2), 135–145.
<https://doi.org/10.1016/j.learninstruc.2007.01.011>
- Jerrim, J., Micklewright, J., Heine, J., Salzer, C. & McKeown, C. (2018). PISA 2015: How big is the ‘mode effect’ and what has been done about it? *Oxford Review of Education*, 44(4), 476-493. doi: 10.1080/03054985.2018.1430025
- Johnson, R.B., Onwuegbuzie, A.J., Turner, L.A. (2007). Towards a definition of mixed methods research. *Journal of Mixed Methods Research*, 112-133.
doi: 10.1177/1558689806298224
- Jordan, K. (1998). Defining multimedia. *IEEE Multimedia*, 5, 8-15.
doi: 10.1109/93.664737.
- Just, M. & Carpenter, P. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 329–355
- Kalyuga, S., & Renkl, A. (2010). Expertise reversal effect and its instructional implications: Introduction to the special issue. *Instructional Science*, 38(3), 209–215.
<https://doi.org/10.1007/s11251-009-9102-0>
- Kanning, U., Grewe, K., Hollenberg, S. & Hadouch, M. (2006). From the subjects’ point of view: Reactions to different types of situational judgment items. *European Journal of Psychological Assessment*, 22(3), 168–176. <https://doi.org/10.1027/1015-5759.22.3.168>
- Karakolidis, A., O’Leary, M. & Scully, D. (2021). Animated videos in assessment: comparing validity evidence from and test-takers' reactions to an animated and a text-based situational judgement test. *International Journal of Testing*, 10.1080/15305058.2021.1916505

- Keane, N. & McInerney, C. (2017). *Report on the provision of Courses in Computer Science in Upper Second Level Education Internationally*. National Council for Curriculum and Assessment.
https://www.ncca.ie/media/2605/computer_science_report_sc.pdf
- Khedher, A. Ben, Jraidi, I., & Frasson, C. (2018). Static and dynamic eye movement metrics for students' performance assessment. *Smart Learning Environments* 5(14), 1-12.
<https://doi.org/10.1186/s40561-018-0065-y>
- Kirschner, P., Park, B., Malone, S. & Jarodzka, H. (2016). Toward a cognitive theory of multimedia assessment. In M.J. Spector, B.B. Lockee & M.D. Childress (Eds.), *Learning, Design, and Technology* (pp. 1-23). Springer.
https://doi.org/10.1007/978-3-319-17727-4_53-1
- Kivunja, C., & Kuyini, A.B. (2017). Understanding and applying research paradigms in educational contexts. *International Journal of Higher Education*, 6(5), 26-41.
<https://doi.org/10.5430/ijhe.v6n5p26>
- Kong, X., Davis, L., McBride, Y. & Morrison, K. (2018). Response time differences between computers and tablets. *Applied Measurement in Education*, 31(1), 17-29.
<https://doi.org/10.1080/08957347.2017.1391261>
- Krstić, K., Šoškić, A., Ković, V., & Holmqvist, K. (2018). All good readers are the same, but every low-skilled reader is different: An eye-tracking study using PISA data. *European Journal of Psychology of Education*, 33, 521-541.
<https://doi.org/10.1007/s10212-018-0382-0>
- Kryptos, A.-M., Klugkist, I., Mertens, G., & Engelhard, I. M. (2019). A step-by-step guide on preregistration and effective data sharing for psychopathology research. *Journal of Abnormal Psychology*, 128(6), 517-527. <https://doi.org/10.1037/abn0000424>
- Kuncel, N.R. & Credé, M. & Thomas, L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research*, 75(1), 63-82.
<https://doi.org/10.3102/00346543075001063>
- Lai, M., Tsai, M., Yang, F., Hsu, C., Liu, T., Lee, S. et al. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review*, 10, 90-115. <https://doi.org/10.1016/j.edurev.2013.10.001>
- Lane, S. (2017). Commentary II. In K. Ercikan & J.W. Pellegrino (Eds.), *Validation of Score Meaning for the Next Generation of Assessments* (pp. 136-147). Routledge.

- Lazonder, A., Walraven, A., Gijlers, H., & Janssen, N. (2020). Longitudinal assessment of digital literacy in children: Findings from a large Dutch single-school study. *Computers & Education*, 143(103681).
<https://doi.org/10.1016/j.compedu.2019.103681>
- Lee, Y. (2018). Effect of uninterrupted time-on-task on students' success in Massive Open Online Courses (MOOCs). *Computers in Human Behavior*, 86, 174–180.
<https://doi.org/10.1016/j.chb.2018.04.043>
- Lee, Y. H., Hao, J., Man, K., & Ou, L. (2019). How do test takers interact with simulation-based tasks? A response-time perspective. *Frontiers in Psychology*, 10, 906.
<https://doi.org/10.3389/fpsyg.2019.00906>
- Lehane, P. (2019). *Leaving Certificate Computer Science: Factors to consider when developing computer-based examinations*. National Council of Curriculum and Assessment.
https://www.ncca.ie/media/4081/lccs_cbe_factorstoconsider_lehane2019-for-ncca-website.pdf
- Leighton, J.P., Gokiert, R.J., Cor, M.K., & Heffernan, C. (2010). Teacher beliefs about the cognitive diagnostic information of classroom-versus large-scale tests: Implications for assessment literacy. *Assessment in Education: Principles, Policy & Practice*, 17, 7–21. doi:10.1080/09695940903565362
- Levy, R. (May, 2012). *Psychometric advances, opportunities, and challenges for simulation-based assessment* [paper presentation]. Invitational Research Symposium on Science Assessment
<https://www.ets.org/Media/Research/pdf/session2-levy-paper-tea2012.pdf>
- Lievens, F. & Sackett, P. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *The Journal of Applied Psychology*, 91(5), 1181–1188. <https://doi.org/10.1037/0021-9010.91.5.1181>
- Lindner, M. (2020). Representations and decorative pictures in science and mathematics tests: Do they make a difference? *Learning and Instruction*, 68, 101345.
<https://doi.org/10.1016/j.learninstruc.2020.101345>
- Lindner, M., Lüdtke, O., Grund, S., & Köller, O. (2017a). The merits of representational pictures in educational assessment: Evidence for cognitive and motivational effects in a time-on-task analysis. *Contemporary Educational Psychology*, 51, 482–492.
<https://doi.org/10.1016/J.CEDPSYCH.2017.09.009>

- Lindner, M., Eitel, A., Strobel, B., & Köller, O. (2017b). Identifying processes underlying the multimedia effect in testing: An eye-movement analysis. *Learning and Instruction, 47*, 91–102. <https://doi.org/10.1016/j.LEARNINSTRUC.2016.10.007>
- Lindner, M., Ihme, J., Saß, S., & Köller, O. (2018). How representational pictures enhance students' performance and test-taking pleasure in low-stakes assessment. *European Journal of Psychological Assessment, 34*(6), 376–385. <http://dx.doi.org/10.1027/1015-5759/a000351>
- Livingston, S. A. (2009). *Constructed response test questions: Why we use them? How we score them?* https://www.ets.org/Media/Research/pdf/RD_Connections11.pdf
- Luke, S. & Asplund, A. (2018). Prereaders' eye movements during shared storybook reading are language-mediated but not predictive. *Visual Cognition, 26*(5), 351–365, doi: 10.1080/13506285.2018.1452323
- MacCann, C., Lievens, F., Libbrecht, N. & Roberts, R. (2016). Differences between multimedia and text-based assessments of emotion management: An exploration with the multimedia emotion management assessment (MEMA). *Cognition and Emotion, 30*(7), 1317–1331. <https://doi.org/10.1080/02699931.2015.1061482>
- Malone, S. & Brünken, R. (2013). Assessment of driving expertise using multiple choice questions including static vs. animated presentation of driving scenarios. *Accident Analysis and Prevention, 51*, 112–119. <https://doi.org/10.1016/j.aap.2012.11.003>
- Martinez, M. (1991). A comparison of multiple-choice and constructed figural response items. *Journal of Educational Measurement, 28*, 131–145. <http://www.jstor.org/stable/1434795>
- Masters, J., & Gushta, M. (2018). *Using technology-enhanced items to measure fourth grade geometry knowledge*. <https://www.measuredprogress.org/wp-content/uploads/2018/05/Using-Technology-Enhanced-Items-to-Measure-Fourth-Grade-Geometry-Knowledge.pdf>
- Maxwell, J. (2005). *Qualitative research design: An interactive approach* (2nd Edition). SAGE Publications.
- Mayer. (2005). *The Cambridge Handbook of Multimedia Learning*. Cambridge University Press.
- Mayer, R. (2008). Applying the science of learning: Evidence based principles for the design of multimedia instruction. *American Psychologist, 63*(8), 760–769. <http://dx.doi.org/10.1037/0003-066X.63.8.760>

- Mayer, R. (2009). *Multimedia Learning* (2nd Edition). Cambridge University Press.
- Mayer, R. (2014). Cognitive theory of multimedia learning. In R.E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (2nd Edition) (pp. 43-71). Cambridge University Press.
- Mayer, R. (2017). Using multimedia for e-learning. *Journal of Computer Assisted Learning*, 33, 403–423. <https://doi.org/10.1111/jcal.12197>
- Mayer, R., & Moreno, R. (2002). Animation as an aid to multimedia learning. *Educational Psychology Review*, 14(1), 87-99. <https://doi.org/10.1023/A:1013184611077>
- Mayrath, M., Clarke-Midura, J., & Robinson, D. (2012). *Technology Based Assessment for 21st Century Skills: Theoretical and Practical Implications from Modern Research*. Springer-Verlag.
- Measured Progress/ETS Collaborative. (2012). *Smarter balanced assessment consortium: Technology Enhanced Items*. <https://www.measuredprogress.org/wp-content/uploads/2015/08/SBAC-Technology-Enhanced-Items-Guidelines.pdf>
- Merriam, S. B., Johnson-Bailey, J., Lee, M., Kee, Y., Ntseane, G. & Muhamad, M. (2005). Power and positionality: Negotiating insider/outsider status within and across cultures. *International Journal of Lifelong Education*, 20(5), 405-416. <https://doi.org/10.1080/02601370120490>
- Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 33–45). Lawrence Erlbaum Associates.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13-23. <https://doi.org/10.3102/0013189X023002013>
- Mislevy, R., Steinberg, L., & Almond, R. (1999). *On the roles of task model variables in assessment design* [paper presentation]. Cognitive Tests: Theory and Practice, New Jersey. <https://cresst.org/wp-content/uploads/TECH500.pdf>
- Mislevy, R., Almond, R. & Lukas, J. (2003). A brief introduction to evidence-centered design. *ETS Research Report RR-03-16*. Educational Testing Service. doi: 10.1002/j.2333-8504.2003.tb01908.x

- Molich, R. & Nielson, J. (1990). Improving a human-computer dialogue: What designers know about traditional interface design. *Communications of the ACM* 33 (March). <https://pdfs.semanticscholar.org/8e67/d5075db82691aad39743d3414019ab4e38c0.pdf>
- Moon, J., Keehner, M., & Katz, I. (2019). Affordances of item formats and their effects on test-taker cognition under uncertainty. *Educational Measurement: Issues and Practice*, 38(1), 54–62. <https://doi.org/10.1111/emip.12229>
- National Council of Curriculum and Assessment (NCCA). (2007). *Assessment in the primary school curriculum: Guidelines for schools*. NCCA. <https://www.ncca.ie/media/1351/assessment-guidelines.pdf>
- Navarro, O., Molina, A., Lacruz, M. & Ortega, M. (2015). Evaluation of multimedia educational materials using eye tracking. *Procedia-Social and Behavioral Sciences*, 197, p. 5-17. <https://doi.org/10.1016/j.sbspro.2015.07.366>
- New Zealand Qualifications Authority (NZQA). (2014). *Report on the eMCAT project*. <https://www.nzqa.govt.nz/assets/About-us/Our-role/innovation/2014-eMCAT-report-final.pdf>
- New Zealand Qualifications Authority (NZQA). (2018). *Exam centre manager/supervisor survey: Digital pilots 2017*. <https://www.nzqa.govt.nz/assets/About-us/Future-State/2017-trials-and-pilots/ECMorSupervisorsurveyanalysisfinal.pdf>
- Nielsen, J. (2012). *Thinking aloud: The #1 usability tool*. <http://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/>
- Nuthmann, A. (2017). Fixation durations in scene viewing: Modelling the effects of local image features, oculomotor parameters, and task. *Psychonomic Bulletin Review*, 24, 370–392. <https://doi.org/10.3758/s13423-016-1124-4>
- Nyström, M., Andersson, R., Holmqvist, K., & van de Weijer (2012). The influence of calibration method and eye physiology on eye tracking data quality. *Behavioural Research*, 45, 272–288. <https://doi.org/10.3758/s13428-012-0247-4>
- O’Leary, M., Scully, D., Karakolidis, A., & Pitsia, V. (2018). The state-of-the-art in digital technology-based assessment. *European Journal of Education*, 53(2), 160–175. <https://doi.org/10.1111/ejed.12271>
- Olsen, A., Smolentzov, L., & Strandvall, T. (September, 2010). Comparing different eye tracking cues when using the retrospective think aloud method in usability testing. *Proceedings of Journal of Usability Studies* (pp. 45-53). British Computer Society. <https://www.scienceopen.com/hosted-document?doi=10.14236/ewic/HCI2010.8>

- Onwuegbuzie, A. & Leech, L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methods. *International Journal of Social Research Methodology*, 8(5), 375-387. <https://doi.org/10.1080/13645570500402447>
- Oranje, A., Gorin, J., Jia, Y., & Kerr, D. (2017). Collecting, analysing, and interpreting response time, eye tracking and log data. In K. Ercikan & J. W. Pellegrino (Eds.), *Validation of Score Meaning for the Next Generation of Assessments* (pp. 39–51). Routledge.
- Organisation for Economic Co-Operation and Development (OECD). (2013). *PISA 2015: Draft Science Framework*. OECD.
- Organisation for Economic Co-Operation and Development (OECD). (2015). *PISA 2015 released field trial cognitive items*. <https://www.oecd.org/pisa/test/PISA2015-Released-FT-Cognitive-Items.pdf>
- Organisation for Economic Co-Operation and Development (OECD). (2016a). *PISA 2015 results: Policies and practices for successful schools* (Volume I). OECD Publishing. <https://www.oecd.org/publications/pisa-2015-results-volume-i-9789264266490-en.htm>
- Organisation for Economic Co-Operation and Development (OECD). (2016b). *Programme for International Student Assessment (PISA) 2015: Assessment and analytical framework*. OECD Publishing. <https://read.oecd.org/10.1787/9789264281820-en>
- Organisation for Economic Co-Operation and Development (OECD). (2016c). *PISA 2015 results: Policies and practices for successful schools* (Volume II). OECD Publishing. <https://www.oecd.org/publications/pisa-2015-results-volume-ii-9789264267510-en.htm>
- Organisation for Economic Co-Operation and Development (OECD). (2017). *PISA: 2015 Technical Report*. OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/PISA2015_TechRep_Final.pdf
- Organisation for Economic Co-Operation and Development (OECD). (2018). *PISA: Programme for International Student Assessment* [webpage]. <https://www.oecd.org/pisa/>
- Organisation for Economic Co-Operation and Development (OECD). (2019). *PISA 2018: Assessment and Analytical Framework*. OECD. <https://doi.org/10.1787/b25efab8-en>

- Organisation for Economic Co-Operation and Development (OECD). (2020). *PISA: Programme for International Student Assessment - Try the Test!* [webpage].
<https://tinyurl.com/2x2wp6kr>
- Orquin, J.L., Ashby, N., & Clarke, A. (2016). Areas of interest as a signal detection problem in behavioural eye-tracking research. *Journal of Behavioural Decision Making*, 29, 103-115. <https://doi.org/10.1002/bdm.1867>
- Orquin, J.L. & Hölmqvist, K. (2018). Threats to the validity of eye-movement research in psychology. *Behavioural Research Methods*, 50(4), 1645–1656.
<https://doi.org/10.3758/s13428-017-0998-z>
- Paivio, A. (1986). *Mental Representations: A Dual Coding Approach*. Oxford University Press.
- Pallant, J. (2007). *SPSS Survival Manual* (3rd Edition). McGraw-Hill Open University Press.
- Parshall, C. & Harmes, J. (2008). The design of innovative item types: Targeting constructs, selecting innovations, and refining prototypes. *CLEAR Exam Review*, 19(2), 18–25.
https://www.clearhq.org/resources/CLEAR_summer08_4.pdf#page=20.
- Parshall, C., Harmes, J., Davey, T. & Pashley, P. (2010). Innovative items for computerized testing. In van der Linden W. and Glas C. (Eds.) *Elements of Adaptive Testing* (pp. 215-230). Springer.
- Paulson, E. J., & Henry, J. (2002). Does the degrees of reading power assessment reflect the reading process? An eye-movement examination. *Journal of Adolescent and Adult Literacy*, 46, 234–244. <https://www.jstor.org/stable/40017130>
- Pedaste, M., Mäeots, M., Siiman, L.A., de Jong, T., Siswa, A.n., Kamp, E.T., Constantinos, C.M., Zacharia, Z.C., & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review*, 14, 47-61.
<https://doi.org/10.1016/j.edurev.2015.02.003>
- Platt, D., & Turney, D. (2014). Making threshold decisions in child protection: A conceptual analysis. *British Journal of Social Work*, 44(6), 1472–1490.
<https://doi.org/10.1093/bjsw/bct007>
- Ponsford, R., Meiksin, R., Bragg, S., Crichton, J., Emmerson, L., Tancred, T., Tilouche, N., Morgan, G., Gee, P., Young, H., Hadley, A., Campbell, R. & Bonell, C. (2021) Co-production of two whole-school sexual health interventions for English secondary schools: Positive choices and project respect. *Pilot & Feasibility Studies*, 7 (50).

<https://doi.org/10.1186/s40814-020-00752-5>

Prensky, M. (2001). Digital natives, digital immigrants: Part 1. *On the Horizon*, 9(5), 1–6.

Professional Testing. (2018). *The drag and drop – Why, when, and how to use this item type*. <http://www.proftesting.com/blog/2016/06/22/drag-drop-item-type/>

Quellmalz, E., Davenport, J., Timms, M., DeBoer, G., Jordan, K., Huang, C. & Buckley, B. (2013). Next-generation environments for assessing and promoting complex science learning. *Journal of Educational Psychology*, 105(4), 1100–1114. <https://doi.org/10.1037/a0032220>

QSR International. (2020). *NVivo 12* [computer software]. http://www.qsrinternational.com/products_nvivo.aspx.

Rayner, K., White, S.J., Johnson, R., Liversedge, S.P. (2006). Reading words with jumbled letters: There is a cost. *Psychological Science*, 17(3), 192-193. <https://doi.org/10.1111/j.1467-9280.2006.01684.x>

Road Safety Authority (RSA). (September, 2018). *Optimize, improve and evolve assessments with integrated technology-enabled tools* [paper presentation]. European Association of Test Publishers (E-ATP) Conference, Athens.

Resnick, L. & Resnick, D. (1992). Assessing the thinking curriculum: New tools for education reform. In B.R. Gifford & M.O. O'Connor (Eds.), *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction* (pp. 37–75). Kluwer Academic Publishers.

Rupp, A., Gushta, M., Mislevy, R., & Shaffer, D. (2010). Evidence-centered design of epistemic games: Measurement principles for complex learning environments. *Journal of Technology, Learning, and Assessment*, 8(4). <https://ejournals.bc.edu/index.php/jtla/article/download/1623/1467/>

Russell, M. (2016). A framework for examining the utility of technology-enhanced items. *Journal of Applied Testing Technology*, 17(1), 20–32. <http://www.jattjournal.com/index.php/atp/article/view/89189/67798%5Cnhttp://www.jattjournal.com/>

Russell, M., & Airasian, P.W. (2012). *Classroom Assessment: Concepts and Applications* (7th Edition). McGraw-Hill.

Russell, M. & Moncaleano, S. (2019). Educational assessment examining the use and construct fidelity of technology-enhanced items employed by k-12 testing

- programs. *Practical Assessment, Research & Evaluation*, 24(4), 286-304. <https://doi.org/10.1080/10627197.2019.1670055>
- Sagoo, M.G., Vorstenbosch, M., Bazira, P., Ellis, H., Kambouri, M., & Owen, C. (2020). Online assessment of applied anatomy knowledge: The effect of images on medical students' performance. *Anatomy Science Education*, 14, 342-351. doi: 10.1002/ase.1965
- Sanchez, C. A., & Goolsbee, J.Z. (2010). Character size and reading to remember from small displays. *Computers and Education*, 55(3), 1056-1062. doi: 10.1016/j.compedu.2010.05.001.
- Scalise, K. & Gifford, B. (2006). Computer-Based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment*, 4(6). <https://ejournals.bc.edu/index.php/jtla/article/view/1653>
- Schnotz, W. & Baadte, C. (2015). Surface and deep structures in graphics comprehension. *Memory and Cognition*, 43(4), 605-618. doi: 10.3758/s13421-014-0490-2
- Schnotz, W. & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction*, 13, 141-156. [https://doi.org/10.1016/S0959-4752\(02\)00017-8](https://doi.org/10.1016/S0959-4752(02)00017-8)
- Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research & Evaluation*, 22(4). <http://pareonline.net/getvn.asp?v=22&n=4>
- Shorten, A. & Smith, J. (2017). Mixed methods research: Expanding the evidence base. *Evidence-Based Nursing*, 20, 74-75. doi: 10.1136/eb-2017-102699
- Shiel, G., Kelleher, C., McKeown, C. & Denner, S. (2016). *Future ready? The performance of 15-year-olds in Ireland on science, reading literacy and mathematics in PISA 2015*. Educational Research Centre.
- Shute, V., Leighton, J., Jang, E. & Chu, M. (2016). Advances in the science of assessment. *Educational Assessment*, 21(1), 34-59. <https://doi.org/10.1080/10627197.2015.1127752>
- Shute, V., & Ventura, M. (2013). *Stealth Assessment: Measuring and Supporting Learning in Video Games*. MIT Press.
- Sikes, P. (2004). Methodology, procedures and ethical concerns. In C. Opie (Ed.), *Doing*

- Educational Research: A guide for First-Time Researchers* (pp. 15-33). SAGE Publications.
- Sireci, G. & Zenisky, A. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S.M. Downing & T.M Haladyna (Eds.), *Handbook of Test Development*. Lawrence Erlbaum Associates. <http://www.danangtimes.vn/Portals/0/Docs/314154729-0805852654.pdf#page=344>
- Sorden, S.D. (2012). Cognitive theory of multimedia learning. In B. Irby, G. Brown & R. Lara-Alecio (Eds.), *Handbook of Educational Theories* (pp. 155-169). Information Age Publishing. http://sorden.com/portfolio/sorden_draft_multimedia2012.pdf
- State Examination Commission. (SEC). (2020a). *Homepage* [website]. <https://www.examinations.ie/>
- State Examination Commission. (SEC). (2020b). *New Leaving Certificate and Junior Certificate/Cycle grading* [webpage]. <https://www.examinations.ie/?l=en&mc=ca&sc=ma>
- State Examination Commission. (2021). *Leaving Certificate Computer Science* [webpage]. <https://www.examinations.ie/?l=en&mc=ex&sc=cs>
- Statistical Package for Social Sciences (SPSS). (2020). *SPSS 28*. [Computer software].
- Sternberg, R.J. (2009). *Cognitive psychology* (4th International Student Edition). Wadsworth.
- Tai, R., Loehr, J., & Brigham, F. (2006). An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. *International Journal of Research and Method in Education*, 29(2), 185–208. <https://doi.org/10.1080/17437270600891614>
- Teig, N, Scherer, R, Kjærnsli, M. (2020). Identifying patterns of students' performance on simulated inquiry tasks using PISA 2015 log-file data. *Journal of Research in Science Teaching*, 57(9), 1400– 1429. <https://doi.org/10.1002/tea.21657>
- Tobii AB (2020). *Eye tracker data quality test report: Accuracy, precision, and detected gaze under optimal conditions—controlled environment for Tobii Pro Fusion*. <https://www.tobiipro.com/siteassets/tobii-pro/accuracy-and-precision-tests/tobii-pro-fusion-accuracy-and-precision-test-report.pdf>

- Tullis, T., & Albert, B. (2013). *Measuring the User Experience: Collecting, Analysing and Presenting Usability Metrics* (2nd Edition). Elsevier.
<https://doi.org/10.1016/C2011-0-00016-9>
- Tuzinski, K. (2013). Simulations for personnel selection: An introduction. In M. Fetzner & K. Tuzinski (Eds.), *Simulations for Personnel Selection* (pp. 1-16). Springer.
- Von der Malsburg, T., & Angele, B. (2017). False positives and other statistical errors in standard analysis of eye movements in reading. *Journal of Memory and Language*, 94, 119–133. <http://dx.doi.org/10.1016/j.jml.2016.10.003>
- Vorstenbosch, M., Bouter, S., van den Hurk, M., Kooloos, J., Bolhuis, S., & Laan, R. (2014). Exploring the validity of assessment in anatomy: Do images influence cognitive processes used in answering extended matching questions? *Anatomical Sciences Education*, 7(2), 107–116. <https://doi.org/10.1002/ase.1382>
- Wan, L., & Henly, G. A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education*, 25(1), 58–78. <https://doi.org/10.1080/08957347.2012.635507>
- Wang, C., Tsai, M. & Tsai, C. (2016). Multimedia recipe reading: Predicting learning outcomes and diagnosing cooking interest using eye-tracking measures. *Computers in Human Behavior*. <https://doi.org/10.1016/j.chb.2016.03.064>
- Ward, T., Bernier, R., Mukerji, C., Perszyk, D., McPartland, J. C., Johnson, E., ... Perszyk, D. (2013). Face Validity. In *Encyclopedia of Autism Spectrum Disorders* (pp. 1226–1227). Springer. https://doi.org/10.1007/978-1-4419-1698-3_308
- Weick, K.E. (1995). *Sensemaking in Organizations*. Sage Publications.
- Wellington, J., Bathmaker, A., Hunt, C., McCulloch, G. & Sikes, P. (2005). *Succeeding with your Doctorate*. SAGE Publications.
- Woo, A., Kim, D., & Qian, H. (2014). *Exploring the psychometric properties of innovative items in CAT*. National Council of State Boards of Nursing [paper presentations]. MARCES Conference, Maryland.
<https://marces.org/conference/InnovativeAssessment/5Woo.pdf>
- Wu, H., Chang, C., Chen, C.-L. D., Yeh, T.K. & Liu, C.C. (2010). Comparison of earth science achievement between animation-based and graphic-based testing designs. *Research in Science Education*, 40, 639–673. <https://doi.org/10.1007/s11165-009-9138-9>

- Wu, H., Kuo, C.-Y., Jen, T.-H., & Hsu, Y.-S. (2015). What makes an item more difficult? Effects of modality and type of visual information in a computer-based assessment of scientific inquiry abilities. *Computers & Education*, 85, 35–48.
<https://doi.org/10.1016/J.COMPEDU.2015.01.007>
- Wise, S. (2018). Controlling construct-irrelevant factors through computer-based testing: disengagement, anxiety, & cheating. *Education Inquiry*, 10(1), 21-33.
<https://doi.org/10.1080/20004508.2018.1490127>
- Wright, C., & Reeves, P. (2016). RadBench: Benchmarking image interpretation skills. *Radiography*, 22(2), 131–136. <https://doi.org/10.1016/j.radi.2015.12.010>
- Zagerman, J., Pfeil, U., Reiterer, H. (October, 2016). *Measuring cognitive load using eye tracking technology in visual computing* [paper presentation]. BELIV '16 (p. 78-89). DOI: <http://dx.doi.org/10.1145/2993901.2993908>
- Zhegallo, A.V., & Marmalyuk, P.A. (2015). ETRAN – R extension package for eye-tracking results analysis. *Perception*, 44(8), 1129-1135.
<https://doi.org/10.1177/0301006615594944>
- Zu, T., Hutson, J., Loschky, L., & Rebello, N. (2018). *Use of eye-tracking technology to investigate cognitive load theory* [paper presentation]. Physics Education Research Conference, Ohio. <https://doi.org/10.1119/perc.2017.pr.113>

Appendix A

Assumptions of the Cognitive Theory of Multimedia Learning

Three main assumptions derived from the field of cognitive psychology underpin the Cognitive Theory of Multimedia Learning (CTML; Mayer, 2009). They are the dual-channels, the limited capacity and the active processing assumptions.

A1. Dual Channel Assumption

The dual channel assumption asserts that humans possess separate channels for processing visual and auditory information (Mayer, 2009). The dual-coding theories associated with this assumption were first conceptualised by Paivio (1986) and Baddeley (1992). They are relatively similar but do differ in their conceptualisations. One is based on presentation mode and the other is based on sensory modality. The presentation-mode approach 'focuses on whether the presented stimulus is verbal in nature (such as spoken or printed words) or non- verbal (such as pictures, video, animation, or background sounds)' (Mayer, 2005, p. 34). One channel processes verbal material and the other channel processes pictorial material and nonverbal sounds. This conceptualisation is most consistent with Paivio's (1986) theory. The sensory-modality approach is consistent with Baddeley's (1992) research involving working memory where one channel 'processes visually represented material and the other channel processes auditorily presented material' (Mayer, 2005, p. 34). Interestingly, a key feature of multimedia learning materials – on-screen text – is treated differently by each of these theories. Text is processed in the verbal channel in the presentation-mode approach associated with Paivio (1986) but in the visual channel in Baddeley's (1992) sensory-modality approach.

Mayer (2009) combined both of these approaches when explaining the dual-channel assumption of the CTML. Mayer applied the sensory-modality approach to distinguish between visual (e.g. pictures, animations, video, and on-screen text) and auditory learning material (e.g. narration). However, he claimed that the presentation-mode approach can be used to support the construction of pictorially based and verbally based models in working memory. Although information enters the information system through one channel, learners, depending on their expertise and overall memory capacity, can convert the representation for processing in the other channel. For example, on-

screen text may initially be processed in the visual channel according to the sensory-modality approach because it is presented to the eyes, but an experienced reader could automatically convert images into sounds, which are then processed through the auditory channel. This cross-channel representation of the same stimulus plays an important role in Paivio's (1986) dual-coding theory and is a key feature of Mayer's dual channel assumption. However, there is a limit on the amount of information that they can process.

A2. Limited Capacity Assumption

Underlying the second assumption of the CTML, and many other cognitive theories of learning, is that the cognitive systems involved in the processing of information are limited in their capacity. As a result, learning can be hindered when cognitive overload occurs and working memory capacity is exceeded. This assumption is draws heavily on Sweller's (Chandler & Sweller, 1991) Cognitive Load Theory. In this case, learners may experience three types of cognitive load as described by De Jong (2010). Intrinsic load refers to the complexity of the information presented, specifically how much must the learner understand. This form of cognitive load is closely related to the expertise of the learner and is considered difficult to manipulate. Extraneous load is derived from the instructional material itself and how much unnecessary processing of irrelevant or unrelated information will the learner have to engage in. Finally, germane load describes the mental effort invested by the learner and involves processes such as selecting, interpreting and organizing (Table A2).

Table A1 Types of Cognitive Load (Chandler & Sweller, 1991; Mayer, 2005)

Load Type	Description
Intrinsic	An unavoidable component of learning new material that is mediated by the complexity of material and the learner's previous knowledge.
Extraneous	Learner must process unnecessary information that does not support learning.
Germane	Learner mentally organises the material and relates it to prior knowledge to create an appropriate mental model.

Based on these definitions, Sorden (2012) argued that extrinsic and germane cognitive load can be controlled by the design and presentation of the materials. This argument has important implications for information acquisition when designing instructional materials. Mayer (2014, p. 36) acknowledged that the constraints on an individuals' processing capacity often force people to make decisions about what 'information to pay attention to, the degree to which we should build connections among the selected pieces of information, and the degree to which we should build connections between selected pieces of information and our existing knowledge'. As a result, learning materials need to be carefully designed so that cognitive load is minimised and learners can make appropriate decisions on what information to attend to so as to maximise their knowledge acquisition.

A3. Active Processing Assumption

The final assumption of the CTML takes the learner's limited capacity into consideration and argues that learning materials require people to actively engage in a range of cognitive processes. These processes include the selection of relevant information and the organisation of this information in a way that will allow its integration with the learner's previous knowledge in the area. This assumption is referred to as 'active processing' and occurs in a learner's working memory. Mayer claimed that individuals engaging with multimedia materials must select only the *relevant* verbal and pictorial information of a lesson or topic and then organise that material into a coherent mental model or representation. New models or representations may then need to be integrated with previous knowledge which then contributes to the overall model constructed. When applying this assumption to the design of multimedia materials, Mayer (2009) noted that materials should have an understandable structure, and it should guide the learner in making an appropriate mental model.

Appendix B

Programme for International Student Assessment (PISA)

B1. Programme for International Student Assessment (PISA): An Overview

The Programme for International Student Assessment (PISA) is an international large-scale assessment coordinated by the OECD (Organisation for Economic Co-operation and Development). It aims to provide cross-nationally comparable evidence of student performance on the skills that are considered to be essential for adult life. PISA is not aligned with any particular school curricula or content as it is designed to assess how successful students at the end of compulsory education are at applying their knowledge to real-life situations (OECD, 2018). It is administered every three years to a representative sample of 15-year old students from participating countries which. In 2018, this involved 79 countries and over 540,000 students (OECD, 2018). Student performance in three key domains is measured: mathematics, literacy and scientific literacy. One of the core domains is tested in detail in each cycle. With the alternating schedule of major domains, a thorough analysis of achievement in each of the three core areas is possible every nine years. In 2015, scientific literacy was the major domain (OECD, 2016a). The use of the term *scientific literacy* rather than *science* illustrates the importance that the PISA science assessment places on the application of scientific knowledge in the context of real-world situations.

Between 2000 and 2012, PISA was carried out as a paper-based assessment (PBA). However, in 2015, pupils in the majority of the participating countries instead took the test on a computer. This represented a major change in procedure for PISA. The use of computers and other digital devices in international large-scale assessments has several attractions as noted by Jerrim et al. (2018). It allows for more efficient test administration along with the immediate availability of results, more varied and authentic test items as well as student preference for such tests. Although these benefits are significant, serious concerns remain regarding potential mode effects, where questions may be systematically harder or easier if they are delivered on paper or on a computer (Jerrim et al., 2018). For example, the average OECD score in PISA 2015 was approximately eight points lower in science than in 2012 (OECD, 2016b). Hong Kong fell by 32 PISA points in science between 2012 and 2015. Science scores decreased by 15 points in Germany and Ireland but they improved by 10 points in Sweden. Using the field

trial data on CBAs and PBAs for Germany, Sweden, and Ireland, Jerrim et al. (2018) aimed to investigate the impact of the two test modes on pupils' responses to reading, science, and mathematics questions originally designed for administration on paper. Mode effects were established in all three cognitive domains. Jerrim et al. (2018) argue that if these are not accounted for, mode effects could limit the comparability of PISA 2015 scores with previous cycles.

While these findings are concerning, PISA's use of technology for testing will continue to expand and will most likely become the dominant mode of administration in the future. As a result, it is essential that the design of items in PISA and other digital assessments be carefully done in accordance with good practice. As a result, exploring the design of items from PISA's TBAs was considered to be particularly useful to the field and worthy of further investigation.

B2. Understanding PISA's Scientific Literacy Framework

Scientific literacy is a widely used term that relates to how an individual uses their scientific skills and knowledge to participate in society (Chadwick et al., 2016). The PISA 2015 Framework for Scientific Literacy (OECD, 2013, p. 7) defined scientific literacy as 'the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen'. The OECD (2019) asserts that an individual's level of scientific literacy is underpinned by three types of scientific knowledge: *content knowledge of science*, *procedural knowledge about science* and *epistemic knowledge about science* (i.e. an understanding of the rationale for the common practices of scientific enquiry, and the meaning of foundational terms). By the time students leave compulsory science education (most commonly at the age of 15), it is assumed that these complementary types of scientific knowledge should have supported their ability to perform the following scientific competencies (OECD, 2019):

- *Explain phenomena scientifically* which involves recognising, offering and evaluating explanations for a range of natural and technological phenomena.
- *Evaluate and design scientific enquiry* which requires describing and appraising scientific investigations and proposing ways of addressing questions scientifically.
- *Interpret data and evidence scientifically* which includes the analysis and evaluation of any data, claims and arguments in a variety of representations, and drawing appropriate scientific conclusions.

Although these competencies do not map directly onto Anderson and Krathwohl's (2001) taxonomy, it appears from the definitions included in PISA 2015 (OECD, 2016a) that the construct of scientific literacy encompasses all levels of this taxonomy. Therefore, use of items adhering to the OECD's (2016a) framework of scientific literacy seems particularly relevant given the importance of developing appropriate assessments that address higher order thinking skills. These competencies are each divided into five sub-competencies, which can be seen in Tables B1, B2 and B3 below (adapted from OECD, 2013, 2019; Chadwick et al., 2016).

Table B1 Explain Phenomena Scientifically

Explain Phenomena Scientifically
Recognise, offer and evaluate explanations for a range of natural and technological phenomena demonstrating the ability to:
A. Recall and apply appropriate scientific knowledge;
B. Identify, use and generate explanatory models and representations;
C. Make and justify appropriate predictions;
D. Offer explanatory hypotheses;
E. Explain the potential implications of scientific knowledge for society.

Table B2 Evaluate and Design Scientific Enquiry

Evaluate and Design Scientific Enquiry
Describe and appraise scientific investigations and propose ways of addressing questions scientifically demonstrating the ability to:
A. Identify the question explored in a given scientific study;
B. Distinguish questions that are possible to investigate scientifically;
C. Propose a way of exploring a given question scientifically;
D. Evaluate ways of exploring a given question scientifically;
E. Describe and evaluate a range of ways that scientists use to ensure the reliability of data and the objectivity and generalisability of explanations.

Table B3 Interpret Data and Evidence Scientifically

Interpret Data and Evidence Scientifically
Analyse and evaluate scientific data, claims and arguments in a variety of representations and draw appropriate conclusions demonstrating the ability to:
A. Transform data from one representation to another;
B. Analyse and interpret data and draw appropriate conclusions;
C. Identify the assumptions, evidence and reasoning in science-related texts;
D. Distinguish between arguments which are based on scientific evidence and theory and those based on other considerations;
E. Evaluate scientific arguments and evidence from different sources (e.g. newspaper, internet, journals).

B3. Scientific Literacy: Test Items

Figure B1 depicts an example of a technology-based item from PISA's 2015 TBA for scientific literacy.

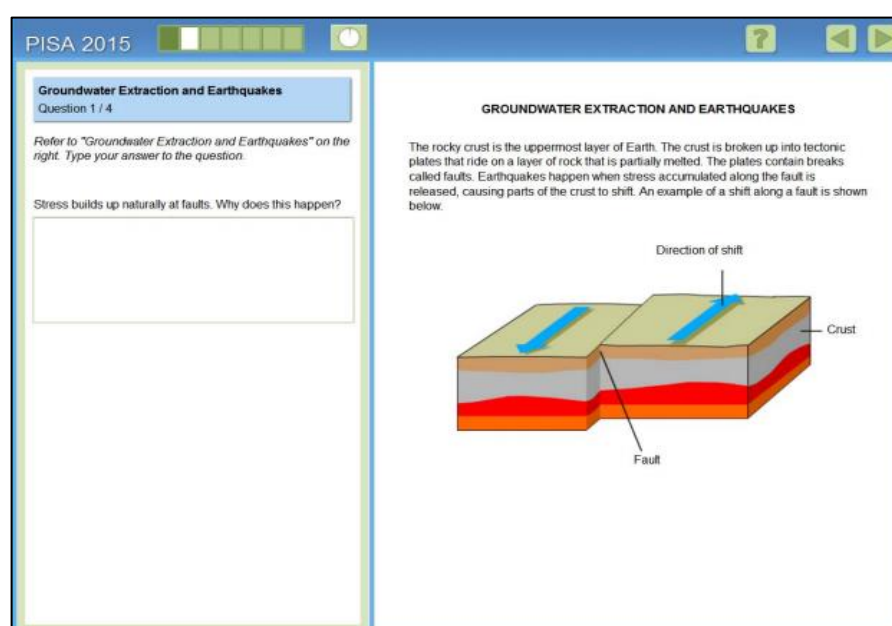


Figure B1 'Groundwater Extraction and Earthquakes' (PISA 2015)

For PISA 2015, each of the items used for the assessment of students' performance in science was mapped against the different aspects of the framework outlined previously,

as well as against three additional dimensions (response format, cognitive demand and context), in order to create a balanced assessment that covers the full framework. Interestingly, the explicit attempt to include items that could cover different levels of cognitive demand across all science competencies and knowledge was a new feature of PISA 2015 (OECD, 2016a). Cognitive demand, according to the OECD (2016a, p. 55) referred to the ‘type of mental processes required to complete an item’.⁶⁵ Items that merely required the recall of one piece of information had low cognitive demands, even if the knowledge needed was quite complex. In contrast, items that required the recall of multiple pieces of knowledge as well as ‘a comparison or evaluation of the competing merits of their relevance would be seen as having high cognitive demand’ (OECD, 2016a, p. 88). The six categories used to classify items have been outlined in Table A4 along with their sub-categories.

Table B4 Categories describing the scientific literacy items constructed for the PISA 2015 Cycle

Competency	Knowledge Types	Content Area	Response Types	Cognitive Demand	Context
Explain Phenomena Scientifically	Content	Physical Systems	Simple Multiple Choice	Low	Personal ⁶⁶
Evaluate and Design Scientific Enquiry	Procedural	Living Systems	Complex Multiple Choice	Medium	Local/ National ⁶⁷
Interpret data and evidence scientifically	Epistemic	Earth and Space Systems	Constructed Response	High	Global ⁶⁸

⁶⁵ The difficulty of any item in PISA 2015, was determined through combining the complexity or range of knowledge required to answer the item with the cognitive operations needed to process the item. Items were classified in PISA 2015 (OECD, 2016a) as low, medium or high in terms of cognitive demand.

⁶⁶ Items related to students’ and families’ daily lives.

⁶⁷ Items involving contexts related to the community in which the student lives.

⁶⁸ Items on issues defined by life across the world.

Appendix C

Items in Static and Dynamic Conditions (Study 1A, Study 1B)

These are adaptations of an original work by the OECD. The opinions expressed and arguments employed in this adaptation are the sole responsibility of the authors of the adaptation and should not be reported as representing the official views of the OECD or of its member countries.

DCU12345

Meteoroids and Craters
Question 1 / 3

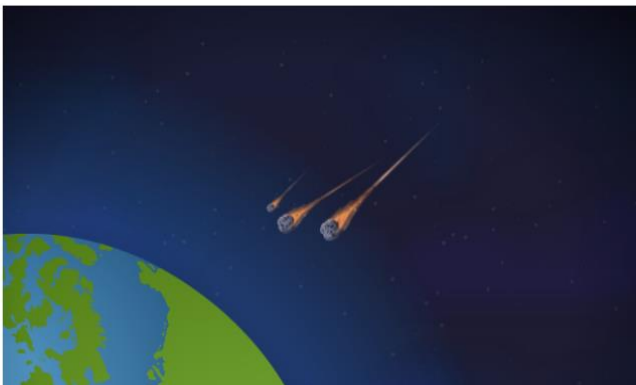
Refer to "Meteoroids and Craters" on the right. Click on a choice to answer the question.

As a meteoroid approaches Earth and its atmosphere, it speeds up. Why does this happen?

☐ The meteoroid is pulled in by the rotation of Earth
☐ The meteoroid is pushed by the light of the sun
☐ The meteoroid is attracted to the mass of the Earth
☐ The meteoroid is repelled by the vacuum of space

METEORIDS AND CRATERS

Rocks in space that enter Earth's atmosphere are called meteoroids. Meteoroids heat up, and glow as they fall through Earth's atmosphere. Most meteoroids burn up before they hit Earth's surface. When a meteoroid hits Earth, it can make a hole called a crater.



DCU12345

Meteoroids and Craters
Question 1 / 3

Refer to "Meteoroids and Craters" on the right. Click on a choice to answer the question.

As a meteoroid approaches Earth and its atmosphere, it speeds up. Why does this happen?

☐ The meteoroid is pulled in by the rotation of Earth
☐ The meteoroid is pushed by the light of the sun
☐ The meteoroid is attracted to the mass of the Earth
☐ The meteoroid is repelled by the vacuum of space

METEORIDS AND CRATERS




Figure C1 Item M1 (Static, Dynamic)

DCU

12345

»

Meteoroids and Craters

Question 2 / 3

Refer to "Meteoroids and Craters" on the right. Select from the drop-down menus to answer the question.

What is the effect of a planet's atmosphere on the number of craters on a planet's surface?

The thicker a planet's atmosphere, the

select

 craters its surface will have because

select


 meteoroids will burn up in the atmosphere.

more

fewer

METEORIDS AND CRATERS

Rocks in space that enter Earth's atmosphere are called meteoroids. Meteoroids heat up, and glow as they fall through Earth's atmosphere. Most meteoroids burn up before they hit Earth's surface. When a meteoroid hits Earth, it can make a hole called a crater.



DCU

12345

»

Meteoroids and Craters

Question 2 / 3

Refer to "Meteoroids and Craters" on the right. Select from the drop-down menus to answer the question.

What is the effect of a planet's atmosphere on the number of craters on a planet's surface?

The thicker a planet's atmosphere, the

select

 craters its surface will have because

select

 m

more

 will burn up in the atmosphere.

more

fewer

METEORIDS AND CRATERS




Figure C2 Item M2 (Static, Dynamic)

DCU

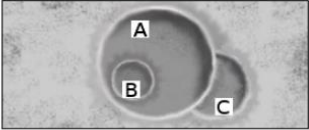
12345

Meteoroids and Craters

Question 3 / 3

Refer to "Meteoroids and Craters" on the right. Use drag and drop to answer the question.

Consider the following three craters.



Put the craters in order by the size of the meteoroids that caused them, from largest to smallest.

Largest

Smallest

A

B

C

Put the craters in order by when they were formed, from oldest to newest.

Oldest

Newest


A

B

C

METEORIDS AND CRATERS

Rocks in space that enter Earth's atmosphere are called meteoroids. Meteoroids heat up, and glow as they fall through Earth's atmosphere. Most meteoroids burn up before they hit Earth's surface. When a meteoroid hits Earth, it can make a hole called a crater.



DCU

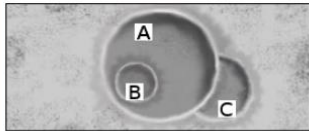
12345

Meteoroids and Craters

Question 3 / 3

Refer to "Meteoroids and Craters" on the right. Use drag and drop to answer the question.

Consider the following three craters.



Put the craters in order by the size of the meteoroids that caused them, from largest to smallest.

Largest

Smallest

A

C

B

Put the craters in order by when they were formed, from oldest to newest.

Oldest

Newest

C

A

B

METEORIDS AND CRATERS

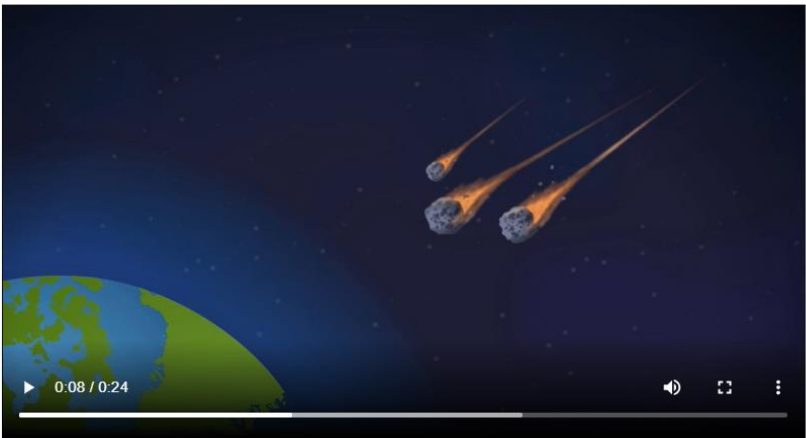


Figure C3 Items M3, M4 (Static, Dynamic)

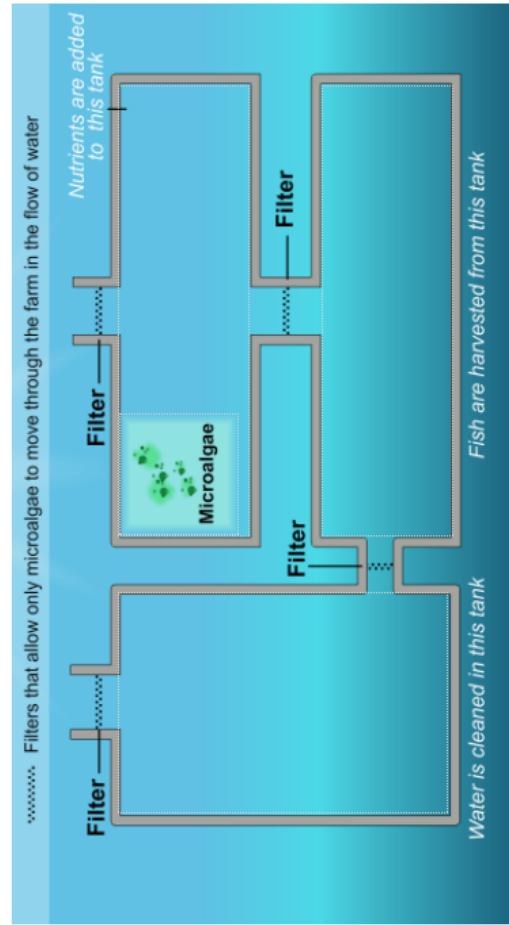
Sustainable Fish Farming

Question 1 / 3

Refer to the information below. Use drag and drop to answer the question.

The diagram shows a design for an experimental fish farm with three large tanks. Filtered salt water is pumped from the ocean before flowing from tank to tank until it is returned to the ocean. The primary goal of the fish farm is to grow common sole to be harvested in a sustainable way.

- Common sole: The fish being farmed. Their preferred food is ragworms
- Microalgae: Microscopic organisms that only need light and nutrients to grow
- Ragworms: Invertebrates that grow very rapidly on a diet of microalgae
- Shellfish: Organisms that feed on microalgae and other small organisms in the water
- Marsh Grass: Grasses that absorb nutrients and wastes from the water



The researchers need to decide in which tank each organism should be placed. Drag and drop each of the organisms below to the appropriate tank above to ensure that the Common Sole is fed and that salt water is returned to the ocean unchanged. The microalgae are already in the correct tank.



Figure C4 Item F1 (Static, Dynamic – Identical items across conditions)

DCU
1
2
3
4
5

Sustainable Fish Farming

Question 2 / 3

Refer to the information below. Click on a choice to answer the question.

The diagram shows a design for an experimental fish farm with three large tanks. Filtered salt water is pumped from the ocean before flowing from tank to tank until it is returned to the ocean. The primary goal of the fish farm is to grow common sole to be harvested in a sustainable way.

- Common sole:** The fish being farmed. Their preferred food is ragworms
- Microalgae:** Microscopic organisms that only need light and nutrients to grow
- Ragworms:** Invertebrates that grow very rapidly on a diet of microalgae
- Shellfish:** Organisms that feed on microalgae and other small organisms in the water
- Marsh Grass:** Grasses that absorb nutrients and wastes from the water

The researchers have noticed that the water that is being returned to the ocean contains a large quantity of nutrients. Adding which of the following to the farm will reduce this problem.

☐ More nutrients
☐ More ragworms
☐ More shellfish
☐ More marsh grass

Figure C5 Item F2 (Static, Dynamic – Identical items across conditions)

DCU
1
2
3
4
5

Sustainable Fish Farming

Question 3 / 3

Refer to the information below. Click on a choice to answer the question.

Which procedure would make fish farming more sustainable?

☐ Increasing the rate of water flow through the tanks.
☐ Increasing the amount of nutrients added to the first tank.
☐ Using filters that allow larger organisms to move between the tanks.
☐ Using the wastes produced by the organisms to make fuel to run the water pumps.

Figure C6 Item F3 (Static, Dynamic – Identical items across conditions)

Blue Power Plant

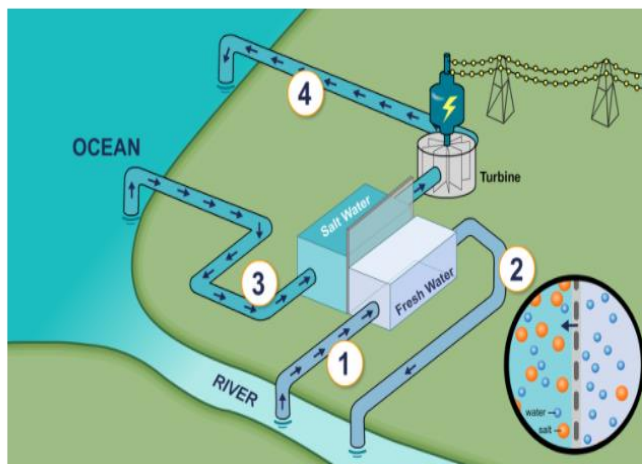
Question 1 / 4

Refer to "Blue Power Plant" on the right. Select one or more boxes.

Four locations in the power plant have been numbered. Water is pumped from the river to location 1, marked on screen.

In which locations could water molecules that come from the river be found later in the process?

- ☐ Location 2
☐ Location 3
☐ Location 4

BLUE POWER PLANT**Blue Power Plant**

Question 1 / 4

Refer to "Blue Power Plant" on the right. Select one or more boxes.

Four locations in the power plant have been numbered. Water is pumped from the river to location 1, marked on screen.

In which locations could water molecules that come from the river be found later in the process?

- ☐ Location 2
☐ Location 3
☐ Location 4

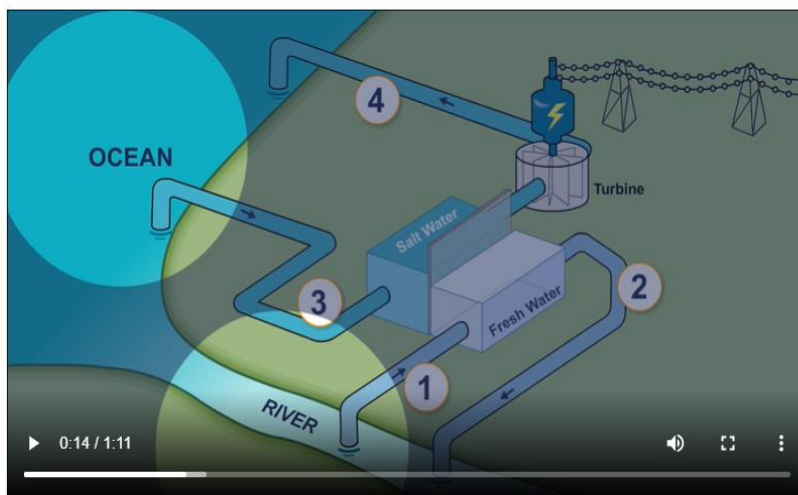
BLUE POWER PLANT

Figure C7 Item P1 (Static, Dynamic)

DCU

1
2
3
4
5

Blue Power Plant
Question 2 / 4

Refer to "Blue Power Plant" on the right. Select from the drop-down menus to answer the question. River water has a low concentration of salt. As the molecules move through the membrane, the salt concentration in the container of fresh water

select

and the salt concentration in the container of salt water

increases
decreases

BLUE POWER PLANT

DCU

1
2
3
4
5

Blue Power Plant
Question 2 / 4

Refer to "Blue Power Plant" on the right. Select from the drop-down menus to answer the question. River water has a low concentration of salt. As the molecules move through the membrane, the salt concentration in the container of fresh water

select

and the salt concentration in the container of salt water

select

BLUE POWER PLANT

Figure C8 Item P2 (Static, Dynamic)

DCU
1
2
3
4
5

Blue Power Plant

Question 3 / 4

Refer to "Blue Power Plant" on the right. Select from the drop-down menus to answer the question.

Several energy conversions occur within the power plant. What kind of energy conversion occurs in the turbine and generator?

The turbine and generator convert

select

to

select

gravitational
potential
kinetic
electrical

BLUE POWER PLANT

DCU
1
2
3
4
5

Blue Power Plant

Question 3 / 4

Refer to "Blue Power Plant" on the right. Select from the drop-down menus to answer the question.

Several energy conversions occur within the power plant. What kind of energy conversion occurs in the turbine and generator?

The turbine and generator convert

select

to

select

gravitational
potential
kinetic
electrical

BLUE POWER PLANT

Figure C9 Item P3 (Static, Dynamic)

DCU

1
2
3
4
5

Blue Power Plant
Question 4 / 4

Refer to "Blue Power Plant" on the right. Type your answer to the question.

Many electric power plants use fossil fuels, such as oil and coal as their energy source.

Why is this new power plant considered to be more environmentally friendly than power plants that use fossil fuels?

BLUE POWER PLANT

DCU

1
2
3
4
5

Blue Power Plant
Question 4 / 4

Refer to "Blue Power Plant" on the right. Type your answer to the question.

Many electric power plants use fossil fuels, such as oil and coal as their energy source.

Why is this new power plant considered to be more environmentally friendly than power plants that use fossil fuels?

BLUE POWER PLANT

1:07 / 1:11

Figure C10 Item P4 (Static, Dynamic)

DCU

1
2
3
4
5

➤

Groundwater Extraction

Question 1 / 4

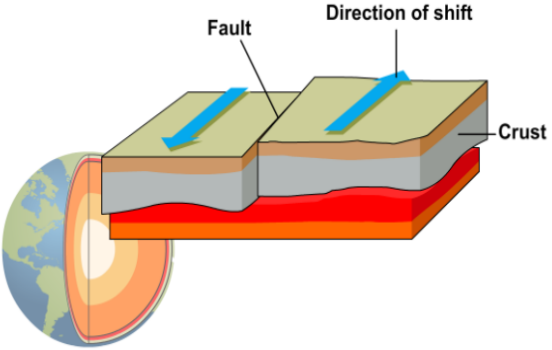
Refer to "Groundwater Extraction" on the right.

Type your answer to the question.

Stress builds up naturally at faults. Why does this happen?

GROUNDWATER EXTRACTION AND EARTHQUAKES

The rocky crust is the uppermost layer of Earth. The crust is broken up into tectonic plates that ride on a layer of rock that is partially melted. The plates contain breaks called faults. Earthquakes happen when stress accumulated along the fault is released, causing parts of the crust to shift. An example of this shift along a fault is shown below.



DCU

1
2
3
4
5

➤

Groundwater Extraction

Question 1 / 4

Refer to "Groundwater Extraction" on the right.

Type your answer to the question.

Stress builds up naturally at faults. Why does this happen?

GROUNDWATER EXTRACTION AND EARTHQUAKES

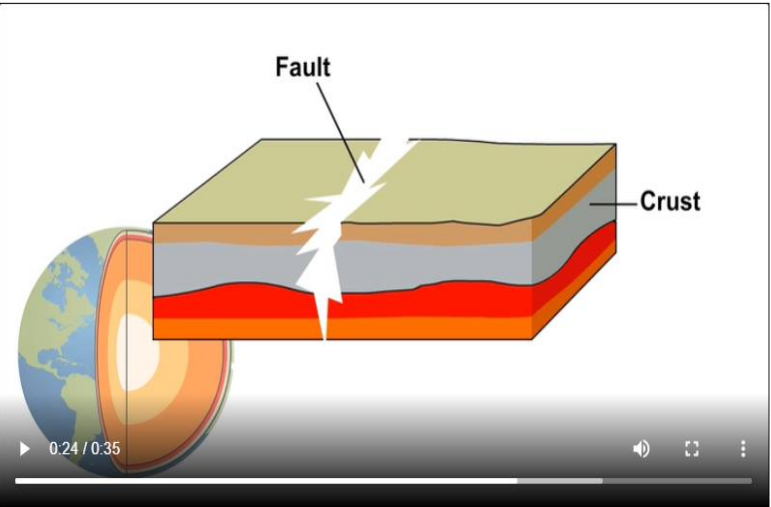


Figure C11 Item G1 (Static, Dynamic)

DCU
1
2
3
4
5

Groundwater Extraction

Question 3 / 4

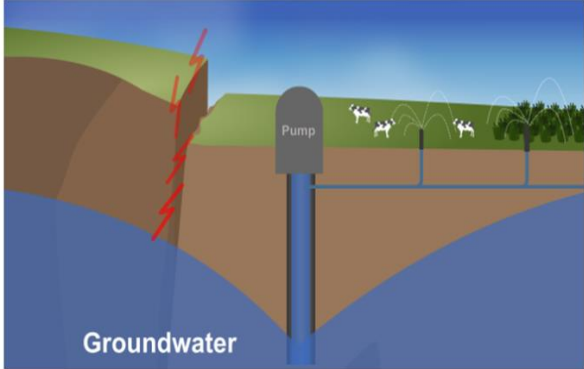
Refer to "Groundwater Extraction" on the right.

Which observation supports the geologists' hypothesis?

- ☐ The earthquake was felt many kilometres away from Lorca.
- ☐ Movement along the fault was greatest in areas where the pumping created the greatest stress.
- ☐ Lorca has had earthquakes that were of greater magnitude than the May 2011 earthquake.
- ☐ The earthquake was followed by a number of smaller earthquakes felt in the region around Lorca.

GROUNDWATER EXTRACTION AND EARTHQUAKES

Lorca, Spain, is located in a region that experiences earthquakes relatively often. One earthquake occurred in Lorca in May 2011. Geologists believe that unlike previous earthquakes in the region, this earthquake may have been caused in part by human activity, specifically by the pumping of groundwater. According to the geologists' hypothesis, extracting water from underground contributed to stress on a nearby fault, which triggered a shift that resulted in the earthquake.



DCU
1
2
3
4
5

Groundwater Extraction

Question 3 / 4

Refer to "Groundwater Extraction" on the right.

Which observation supports the geologists' hypothesis?

- ☐ The earthquake was felt many kilometres away from Lorca.
- ☐ Movement along the fault was greatest in areas where the pumping created the greatest stress.
- ☐ Lorca has had earthquakes that were of greater magnitude than the May 2011 earthquake.
- ☐ The earthquake was followed by a number of smaller earthquakes felt in the region around Lorca.

GROUNDWATER EXTRACTION AND EARTHQUAKES

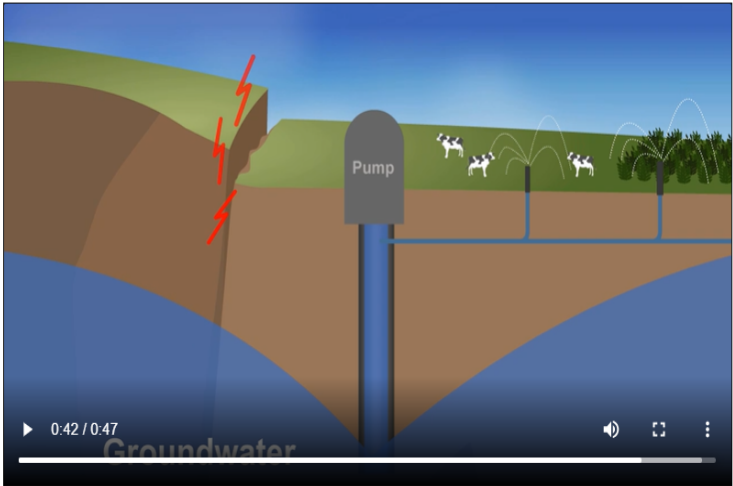


Figure C13 Item G3 (Static, Dynamic)

DCU
1 2 3 4 5

Groundwater Extraction

Question 4 / 4

Refer to "Groundwater Extraction" on the right.

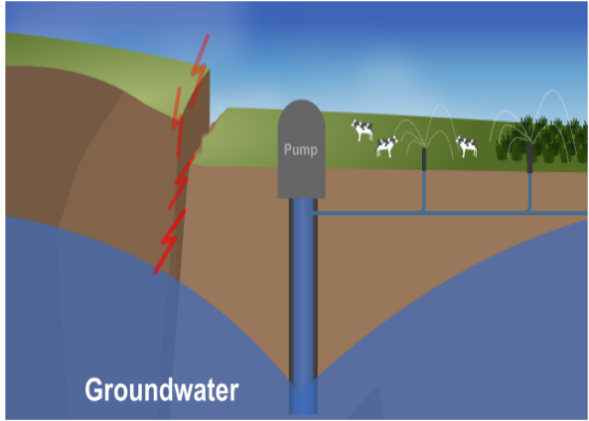
A student who lives in a town in a region far from Lorca learns about the geologists' hypothesis about the 2011 earthquake in Lorca. The student knows that groundwater extraction in the region where he lives has led to a decline in the groundwater level. He is concerned about the possibility of earthquakes in his town. Which of the following questions should the student consider in evaluating the risk that groundwater extraction will trigger an earthquake in his town?

Remember to select **one or more** boxes

- ☐ Does the crust in the region contain faults?
- ☐ Is the crust in the region subject to stress from natural causes?
- ☐ Is water pumped from the ground in the region polluted?
- ☐ What are the average daily temperatures in the region?

GROUNDWATER EXTRACTION AND EARTHQUAKES

Lorca, Spain, is located in a region that experiences earthquakes relatively often. One earthquake occurred in Lorca in May 2011. Geologists believe that unlike previous earthquakes in the region, this earthquake may have been caused in part by human activity, specifically by the pumping of groundwater. According to the geologists' hypothesis, extracting water from underground contributed to stress on a nearby fault, which triggered a shift that resulted in the earthquake.



DCU
1 2 3 4 5

Groundwater Extraction

Question 4 / 4

Refer to "Groundwater Extraction" on the right.

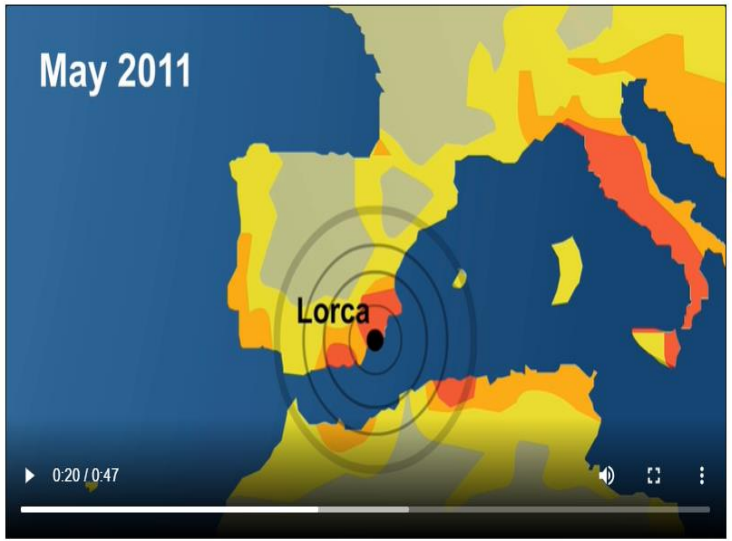
A student who lives in a town in a region far from Lorca learns about the geologists' hypothesis about the 2011 earthquake in Lorca. The student knows that groundwater extraction in the region where he lives has led to a decline in the groundwater level. He is concerned about the possibility of earthquakes in his town. Which of the following questions should the student consider in evaluating the risk that groundwater extraction will trigger an earthquake in his town?

Remember to select **one or more** boxes

- ☐ Does the crust in the region contain faults?
- ☐ Is the crust in the region subject to stress from natural causes?
- ☐ Is water pumped from the ground in the region polluted?
- ☐ What are the average daily temperatures in the region?

GROUNDWATER EXTRACTION AND EARTHQUAKES

May 2011



0:20 / 0:47

Figure C14 Item G4 (Static, Dynamic)

Appendix D

Introductory Screens of Testing Platform

DCU

INTRODUCTION (1/2)

Hello!

My name is Paula Lehane and I am conducting doctoral research at Dublin City University (DCU). I am researching the best ways to design online exams and assessments. This research will be very important in understanding how the use of different types of test questions (e.g. questions that do or don't have pictures or animations, multiple choice/ short answer questions etc.) can influence the scores and test-taking behaviours of secondary school students.

If you agree to take part, you will be asked to complete a short assessment and survey online. This will be good practice for you if you will be doing any online tests or exams in the coming weeks.

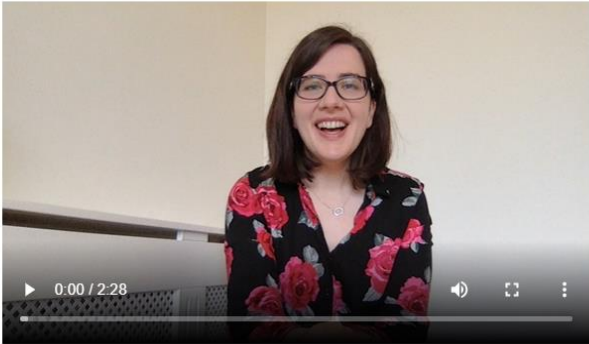
The questions on the assessment will test your everyday **Science** knowledge. You will need to use a pair of **headphones** or have your **audio** on to complete this test. At the end of the test, you will be asked to fill in a short survey (e.g. age, Junior Cert results etc.). Please use a computer, laptop or tablet to complete this test. Phones **can not** be used.

You will **not** need to study anything at all to complete this test. The entire process will take **30-40 minutes**. It is really, really important that you do not try to find the answers to the test online or in your textbooks – your teachers won't see your test results and your information will be completely anonymous. By doing this online exam properly, you can help me to understand how to design better online exams – information that could be very useful in the future.

When I am finished my research, I will let you know through your teachers about what I found and they will pass this information on to you.

I understand what is involved in this study and I agree to take part. ☐

I understand the importance of the study and therefore I will avoid looking at textbooks and other online resources while doing this test. ☐



DCU

INTRODUCTION (2/2)

There are **5 Sections** in this exam.


Each section has 2-5 questions for you to answer. There are **16 questions** in total.

For each question, you will need to read a piece of text or watch a video to help you answer the questions. If you must watch a video, remember to have your sound on. You can watch the video as many times as you like.

Once you have answered a question and moved to the next one, you will not be able to go back and change your answer.

The first question is a **practice one**.

Thank you for agreeing to do this – good luck



Go to questionnaire

Appendix E

Study 2 Materials

These items are taken directly from the OECD website (<https://tinyurl.com/33xwr3tj>)

PISA 2015

Running in Hot Weather
Introduction

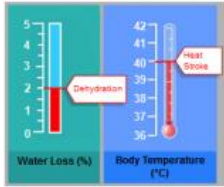
Read the introduction. Then click on the NEXT arrow.

RUNNING IN HOT WEATHER

During long-distance running, body temperature rises and sweating occurs.

If runners do not drink enough to replace the water they lose through sweating, they can experience dehydration. Water loss of 2% of body mass and above is considered to be a state of dehydration. This percentage is labeled on the water loss meter shown below.

If the body temperature rises to 40°C and above, runners can experience a life-threatening condition called heat stroke. This temperature is labeled on the body temperature thermometer shown below.



PISA 2015




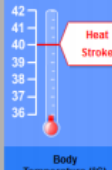
Running in Hot Weather
Introduction

This simulation is based on a model that calculates the volume of sweat, water loss, and body temperature of a runner after a one-hour run.

To see how all the controls in this simulation work, follow these steps:

1. Move the slider for **Air Temperature**.
2. Move the slider for **Air Humidity**.
3. Click on either "Yes" or "No" for **Drinking Water**.
4. Click on the "Run" button to see the results. Notice that a water loss of 2% and above causes dehydration, and that a body temperature of 40°C and above causes heat stroke. The results will also display in the table.

Note: The results shown in the simulation are based on a simplified mathematical model of how the body functions for a particular individual after running for one hour in different conditions.

Air Temperature (°C) 20 25 30 35 40
Air Humidity (%) 20 40 60
Drinking Water ☒ Yes ☐ No

Run

Air Temperature (°C)	Air Humidity (%)	Drinking Water	Sweat Volume (Litres)	Water Loss (%)	Body Temperature (°C)

Figure E1 Explanatory Text/Practice Task

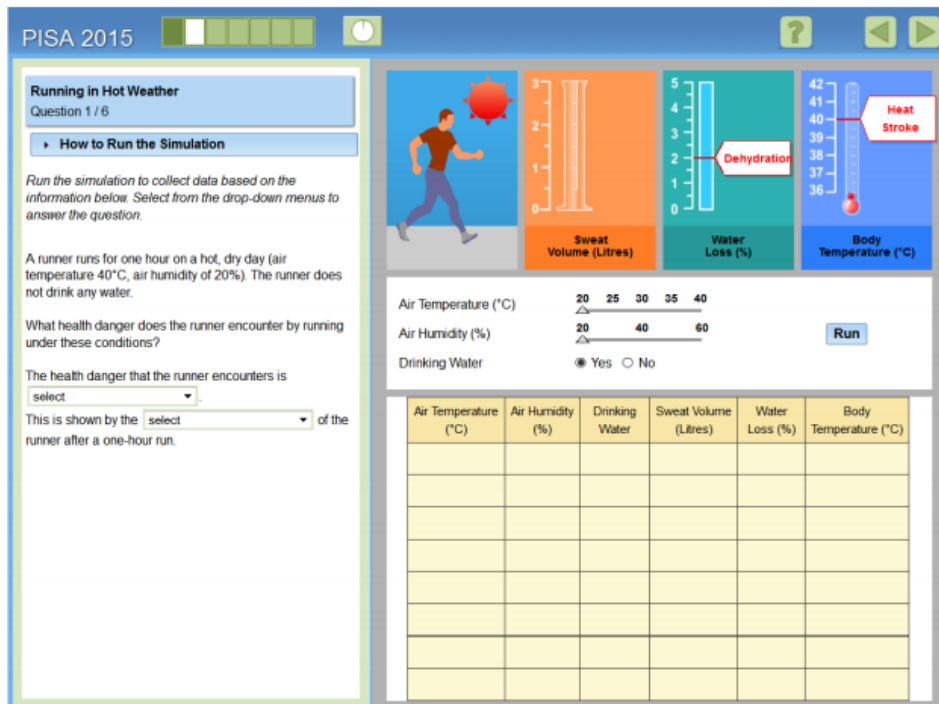


Figure E2 Task 1

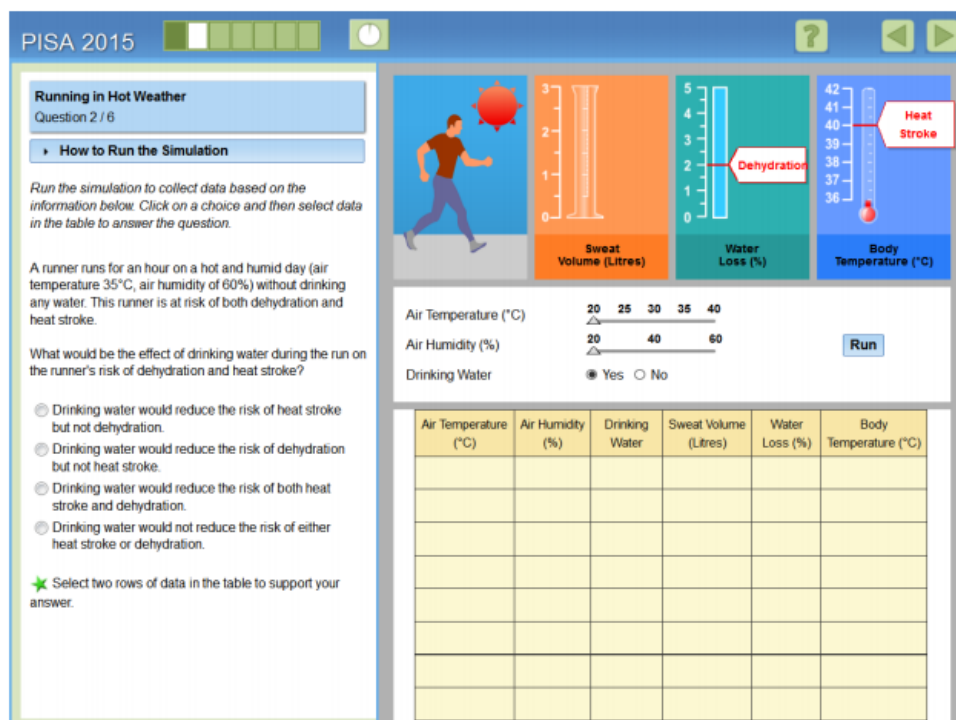


Figure E3 Task 2

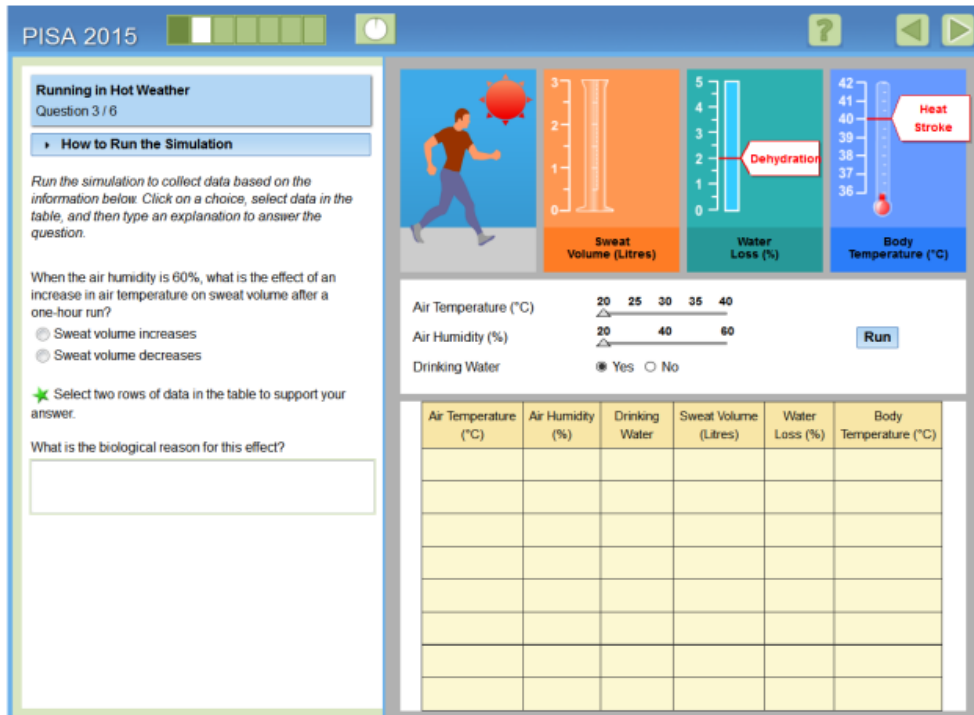


Figure E4 Task 3

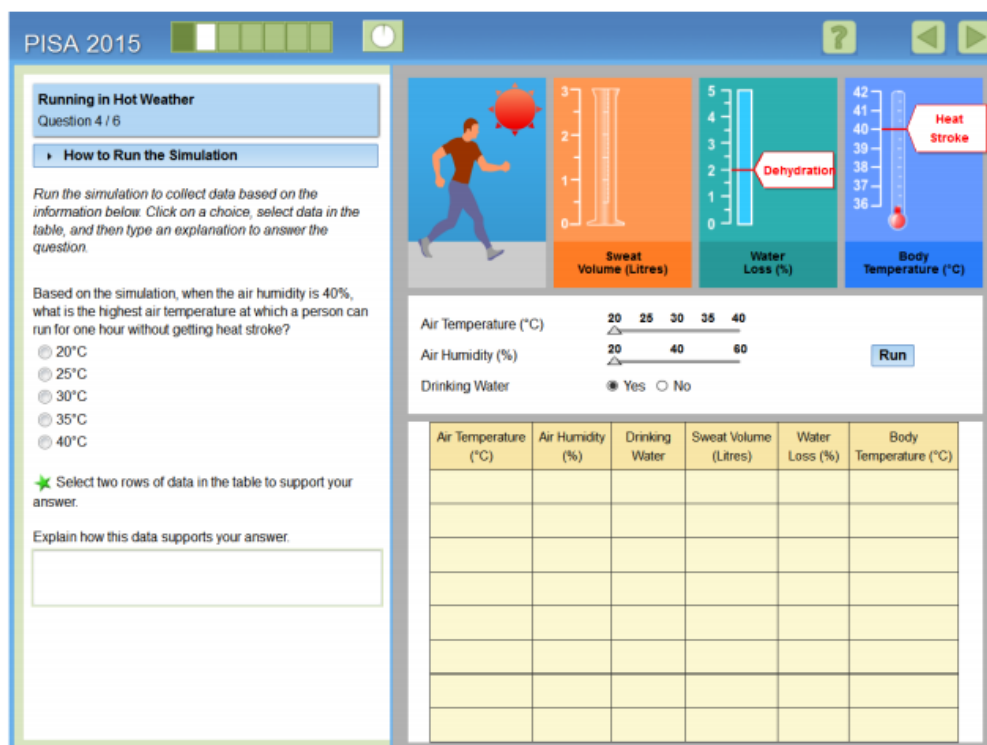


Figure E5 Task 4

Appendix F

Designing an Eye-Tracking Study

F1. Ensuring Data Quality

Recording high quality eye movement data is a requirement for producing replicable results and for drawing well-founded conclusions. The current study had a number of procedures in place to ensure that the data obtained from this study were of high quality. Carter and Luke (2020) noted that the quality of eye movement data derived from any eye tracker can be described in terms of accuracy and precision. Eye tracking data is accurate if the ‘measured eye position corresponds to the actual eye position’, while eye tracking data is precise if it offers ‘consistent measurements of eye position’ (Carter & Luke, 2020, p. 53). Poor accuracy creates an error in determining the true location of the pupil. Poor precision inserts significant ‘noise’ into the data collected as the eye tracker fails to achieve a stable picture of the eye’s pupil. As seen in Figure G1 (where each cross represents the eye’s pupil and each red circle represents a sample collected by the eye tracker), accuracy involves the eye tracker capturing the pupil’s spatial position in relation to a target area whereas precision involves consistent measurements near the pupil. Measures of accuracy and precision were considered when choosing an eye tracker for the current study.

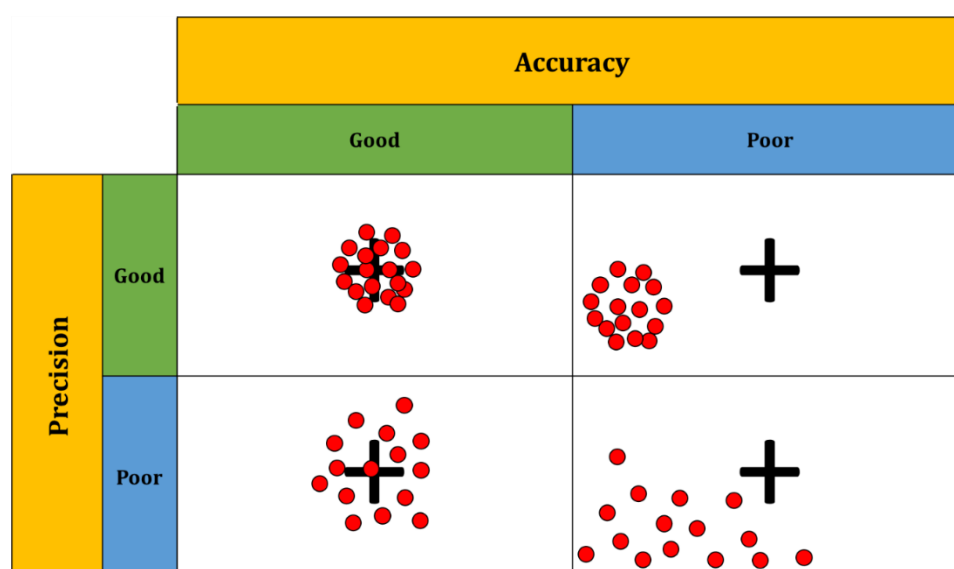


Figure F1 Good vs Poor Eye Movement Accuracy and Precision (adapted from Dalrymple et al., 2018)

There are a wide variety of commercial eye trackers available to researchers (Carter & Luke, 2020). Trackers mainly vary in their speed of data acquisition, as measured in Hertz (Hz), and their set-up e.g. stationary eye trackers, mobile eye trackers. Taking into consideration the funding available. the intended participants of the current study and the data required, a stationary eye tracker was considered to be the most appropriate for the current study. The eye tracker used for the current study was the *Tobii Pro Fusion*, a screen based eye tracker that tracks both eyes while tolerating a variety of head movements and a wide range of physiological variations e.g. eye colour, use of bi-focal glasses, use of contact lenses (see Figure G2). Accuracy and precision test reports on this eye-tracker demonstrated that it can collect highly accurate (within $.3^{\circ}$) and precise (within $.2^{\circ}$) eye movement data (Tobii AB, 2020). Its sampling rate of 120Hz allows for 120 data points for each eye to be captured every second, allowing for a more accurate estimate the true path of the eye when it moves. Although eye trackers with higher sampling rates were available, a lower sampling rate was deemed acceptable as the study was mainly interested in recording where a participant looked, thus negating the need for accuracy beyond milliseconds.



Figure F2 Tobii Pro Fusion

While the accuracy and precision of the data provided by an eye-tracker are usually guaranteed by the manufacturer (e.g. Tobii UK), Holmqvist et al. (2011) cautioned that data quality is also influenced by the experimental setup, the participant, the operator setting up the eye image and the physical recording environment (e.g. lighting).

For the purposes of this study, a number of procedures were developed to ensure and maintain data quality, mainly surround the process of calibration (as outlined in Chapter 3). At the beginning of every eye tracking session, a calibration was conducted to ensure that the eye tracker could capture precise and accurate eye movement data for the individual participant. This was done to establish a 'map' between features detected in the eye image and the physical orientation of the eye (Nyström et al., 2012). Calibration involved the participants looking at nine predefined positions in the stimulus place, as highlighted by a circle. Participants were encouraged to concentrate their gaze on their circle as best they could during the calibration process. At each point, the eye tracker captured a number of eye-image features and associated 'their positions in the eye image with the position of the target' (Nyström et al., 2012, p. 274). The process of calibration was fully automatic and controlled by the system and the calibration procedures applied in the current study have been outlined in Chapter 3.

Appendix G

Interview Schedule: cued-Retrospective Think Aloud (c-RTA)

Explanation: A traditional ‘think-aloud’ asks participants to verbalise their thoughts as they conduct a particular task (Coleman, 2019). Retrospective think aloud is a technique in which users are asked to verbalise their thoughts after performing tasks. In eye-tracking studies, a retrospective think aloud can be ‘cued’ with recordings of a participant’s gaze paths. An eye gaze video shows a video recording of the participant’s interaction with the page on which the eye movements and fixations are also shown (Elbabour et al., 2017).

Script (Indicative Content):

PI (start): Thanks very much for agreeing to talk about your experiences of this test with me. I’m going to show you a slowed down video now of your eye movements for three of the five sets of test questions that you did.

Before I show that to you though, I’m going to show you a short video of my eye-movements. You can see how the screen shows the order of things I looked at with the numbers in the circles. The bigger the circle, the longer I stayed looking at something.

I’m now going to show you your video now. It’ll be at a slowed down pace. While you’re watching the video, tell me what you were thinking if you can remember. If you need to, you can speed up, slow down or pause the video at any time.

PI (end): That was great. Thanks so much for your time.

Notes:

- The PI used acknowledgement tokens to encourage the participants to continue to verbalise their thoughts e.g “aha,” “yeah,” “I see,” and “ok,” as well as reminding participants to think aloud, if needed, using phrases such as “keep talking please” and “please tell me what were you thinking of at this stage.”

Appendix H

Survey

Thank you for completing this test!

Before you go, please fill in the following details:

1. What age are you?

Please Select

2. Think about the last three tests you did in ENGLISH (e.g. Summer tests, Mock Exams, Class Tests). What percentage did you get in each? If you can't remember the exact percentage, just give your best estimate of what you think it was. Please don't leave any space blank.

Test 1 0-100 Test 2 0-100 Test 3 0-100

3. Think about the last three tests you did in MATHS (e.g. Summer tests, Mock Exams, Class Tests). What percentage did you get in each? If you can't remember the exact percentage, just give your best estimate of what you think it was. Please don't leave any space blank.

Test 1 0-100 Test 2 0-100 Test 3 0-100

4. Think about the last three tests you did in SCIENCE (e.g. Summer tests, Mock Exams, Class Tests). What percentage did you get in each? If you can't remember the exact percentage, just give your best estimate of what you think it was. Please don't leave any space blank.

Test 1 0-100 Test 2 0-100 Test 3 0-100

5. To what extent do you agree with the following statements?

☐ I generally have fun when I am learning about different topics in science class.

Please Select

☐ Please Select

☐ Strongly Agree

☐ Agree

☐ Disagree

☐ Strongly Disagree

Please Select

☐ I enjoy acquiring new knowledge in science.

Please Select

☐ I enjoy acquiring new knowledge in science.

Please Select

☐ I am interested in learning about science.

Please Select

6. What is your level of interest in the following Science topics:

☐ Biospheres (e.g. eco-systems, sustainability)

Please Select

☐ Motion and Forces (e.g. velocity, friction, magnetic and gravitational forces)

Please Select

☐ Energy and its transformations (e.g. conservation, chemical reactions)

Please Select

☐ The Universe and its history

Please Select

☐ How science can help us prevent disease

Please Select

7. Did you use any of the following supports during the test? Check all that apply.

☐ Text-to-Speech software

☐ Dictionary

☐ Speech-to-Text software

Other

Submit

Appendix I

Standardised Instructions for Study 1A, Study 1B

Thank you for helping me to find out more about the design and use of digital tests and assessments for secondary school students. I **really** appreciate it.

Preparation: The link to this study is: <https://plp.psycholate.com/>. Have this link open on the interactive whiteboard if possible. When the students have settled into the computer room/ classroom, check that the students have:

- A computer/ tablet in front of them that has a reliable connection to the internet and enabled for sound
- The test link open in their web browser (<https://plp.psycholate.com/>). All other tabs and web browsers should be closed.
- Their own headphones
- Guardian consent (names of children with guardian consent will be provided to you in advance)

Please read out the following to the students:

*Today, you're going to be doing a short activity on the computer to help a researcher in Dublin City University better understand how different types of questions and media in an online test or exam can affect student performance. Paula will explain it here [play video on Page 1 of the test link to the whole class]**

* If students query how they can tell the researcher about their Junior Cert results when they did not do these exams in 2020, tell them not to worry and that this question has been changed to reflect their situation and will just ask them about recent class tests instead.

Ask the students to read the instructions on Page 1 and select the two checkboxes at the end if they agree to do the study. If there are any students in the classroom who do not have guardian consent/ do not agree to do the study, please follow normal school procedures for free periods e.g. doing homework etc., but ask the students to not disturb others doing the test.

Play the video on Page 2 to the whole class, where Paula explains the test. After watching the video, please remind the students that:

- Some questions will be hard and some will be easy. They just need to try their best.
- Some students will have to read a short piece of text; others will have to watch. This is randomly assigned by the computer system.
- Once they submit an answer, they cannot go back.
- If there are any students with hearing impairments, please let Paula know this in advance and she will provide a new link to a slightly different test. *

* If some students need to use other specialised software e.g. speech-to-text or other supports e.g. a reader or scribe, please support them in this. They can let the researcher know about this at the end of the study.

For the Practice Question (Bird Migration), they must select the correct option (which is the **first option**) to progress. Feel free to tell the students the answer if they can't get it correct by themselves.

During the test and survey: Try to not give any hints to students about any answers but give them any encouragement they need to finish the test e.g. '*Have a go*', '*You need to put the boxes in the right order using drag-and-drop*', '*Don't worry about spelling*' '*Fast Forward the video to see the locations mentioned in the question*' etc., When the students are finished, they can fill out the survey. Please give them any clarity they need for this (e.g. students are asked to give the score, or their nearest estimate of it, for their three most recent English, Maths and Science Tests. These can be class tests, summer tests, midterm tests etc.,). If students used a scribe or a reader, they can record this in Q7 of the survey. If they used none of these supports, they can skip this question and hit 'submit'.

After the test: Thank them and tell them that their work will help to design better online exams for Irish secondary school students, which are likely to become more common over the next decade. If they want, they can get the answers from their teacher in class or they can ask for the video link explaining the answers (which is available in a separate document).

Appendix J

DCU Research Ethics Committee Approval

Ollscoil Chathair Bhaile Átha Cliath
Dublin City University



Ms. Paula Lehane
School of Policy and Practice, Institute of Education

Prof. Michael O'Leary
School of Policy and Practice, Institute of Education

16th December 2019

REC Reference: DCUREC/2019/208

Proposal Title: A study on the use of different item types in technology-based assessments and their impact on test-taker attention, behaviour and performance

Applicant(s) Ms. Paula Lehane, Prof. Michael O'Leary, Dr Darina Scully, and Prof. Mark Brown

Dear Colleagues,

Further to expedited review, the DCU Research Ethics Committee approves this research proposal.

Materials used to recruit participants should note that ethical approval for this project has been obtained from the Dublin City University Research Ethics Committee.

Should substantial modifications to the research protocol be required at a later stage, a further amendment submission should be made to the REC.

Yours sincerely,

A handwritten signature in black ink, appearing to read 'Geraldine Scanlon', is written over a light grey rectangular background.

Dr Geraldine Scanlon
Chairperson
DCU Research Ethics Committee



Taighde & Nuálaíocht Tacalocht
Ollscoil Chathair Bhaile Átha Cliath,
Baile Átha Cliath, Éire

Research & Innovation Support
Dublin City University,
Dublin 9, Ireland

T +353 1 700 8000
F +353 1 700 8002
E research@dcu.ie
www.dcu.ie



Correspondence from Dublin City University's Research Ethics Committee approving the initial study in 2019 and the pandemic related amendments that were required for 2020.

Appendix K

Study 3 Analysis: Step 2 (Initial Coding)

Transcript Excerpt: Meteoroids Unit (Item M3 and M4) (Interviewee 6)

Researcher: What are you thinking as you're going down through this?

Code: Confusion

Interviewee 6: I was wondering what I was meant to do at the start.

Code: Familiarity (Lack of)

Researcher: And what did you think of having the A, B and C... Did you understand immediately what to do?

Interviewee 6: I wasn't entirely sure at first.

Code: Confusion

Researcher: Mm hmm. So is that why you kind of held the B and let it go?

Interviewee 6: Yes. Just to see what would happen.

Code: Experimentation

Researcher: OK. And look... Look at what you're doing there. You tell me a little bit about what you're doing there, just kind of watch where you're doing.

Interviewee 6: Checking my answer.

Code: Double Checking responses

Researcher: How are you doing that?

Interviewee 6: Well, you can see how I'm doing it. I'm making some weird triangle shape cos I'm going from the ABC to the picture.

Researcher: Yeah, so you're going up to something in the target picture, then you go down and then you go across and then you go up-down-across, up-down-across. And so...so I just want to pause it there. And what did you think of this drag and drop item?

Interviewee 6: I found it confusing at first.

Code: Confusion

Researcher: OK... You know, there was another drag and drop item with the areas of stress. Were you still confused about these items when you completed that one?

Code: Increased familiarity

Interviewee 6: Yeah, I knew what to do straight away there. I wasn't confused for that one.

Researcher: So you didn't think they were too bad, but they would be confusing at first.

Interviewee 6: Yeah.

Code: Confusion

Researcher: And what made them confusing at first?

Interviewee 6: Yeah, I didn't know whether if they were meant to move or not.

Appendix L

Study 3 Analysis: Step 3 (Theme Search)

Excerpts (sample)	Code(s) (sample)	Theme Label	Description
<p>I was confused and how it was supposed to work, how they're supposed to move (I1)</p> <p>They (Drag-and-Drop) were a bit confusing at the start but I got into it then. (I11)</p> <p>This one (Item M3) confused me. I think it was just because it's the first one. (I4)</p>	<p>Confusion</p> <p>Confusion Familiarity (Lack of) Increased familiarity</p> <p>Confusion Familiarity (Lack of) Increased familiarity</p>	<i>Initial Encounters</i>	This theme describes how participants responded when they first logged onto the TBA or when they encountered a 'new' item type. In recalling these events, they acknowledged their initial confusion and how this confusion declined as they completed other similar items.
<p>I kind of remembered most of that sort of thing, because we're doing a lot of stuff on earthquakes at the moment (in class). (I5)</p> <p>Looking at the question before I read the text of something that I usually do. Then I look for those words (I9)</p>	<p>Prior Knowledge</p> <p>Key words Scanning/ Skimming</p>	<i>Test-Taking Strategies</i>	This theme discusses the test-taking strategies identified by participants as they attempted to complete the test items.

Appendix M

Study 3 Analysis: Step 4/Step 5 (Theme Review/Definition)

Step 3: Theme Search	Step 4: Theme Review	Step 5: Theme Definition
<p><i>Initial encounters</i></p> <p>This theme describes how participants responded when they first logged onto the TBA or when they encountered a ‘new’ item type. In recalling these events, they acknowledge their initial confusion and how this confusion declined as they completed other similar items.</p> <p>Relevant codes:</p> <ul style="list-style-type: none"> • Confusion • Familiarity (Lack of) • Increased familiarity • Navigation (within system) • Navigation (within item) • Looking for/ reading instructions • ‘Getting used to it’ • ‘Figuring it out’ • Scanning/ Skimming 	<p>When reviewing the excerpts organised under the theme of ‘<i>Initial encounters</i>’, it became clear that there was some overlap between this theme (<i>Initial encounters</i>) and a number of others identified in Step 3 (e.g. <i>Test Taking Strategies</i>, <i>Understanding the question</i> etc.). In identifying the distinct elements of this theme, it became clear that participants engaged in a ‘familiarisation’ period when they first logged onto the TBA and when they encountered a new item type. Their behaviours during this time were different to the behaviours they exhibited when they were attempting to understand the test question (see Step 5).</p>	<p><i>Familiarisation</i></p> <p>This theme refers to test-takers’ initial interactions to the online testing environment for Study 1B and Study 2. This theme captures how test-takers became familiar to the online testing environment and the different item types contained therein (e.g. the spatial position of item elements, how to select an answer). It also includes test-taker’s thoughts on this process. The behaviours in this theme are distinct from those behaviours in the ‘<i>Sense-making</i>’ theme as the participants did not seem to cognitively engage with the item content e.g. there was no searching for relevant information etc.</p>