Can Retrieval Practice of *The Testing Effect* Increase Self-efficacy in Tests and Reduce Test Anxiety, in 10- to 11-Year-olds?

UNIVERSITY OF CAMBRIDGE

Faculty of Education



Helen Barsham

Hughes College

Supervisor: Dr. Michelle Ellefson

This dissertation is submitted for the degree of Doctor of Education.

May 2021

Declaration of Originality

I can confirm that this thesis is the result of my own work and includes nothing that is the outcome of work completed in collaboration except as specified in the text. This thesis has not been submitted for any other degree or professional qualification except for the Doctorate in Education. It does not exceed the prescribed word limit for the relevant Degree Committee.

Helen Barsham 6th May 2021

Dissemination

Journals

Barsham, H., & Ellefson, M.R. (2020) Can teaching upper primary about the testing effect increase feelings of confidence about test taking? *Journal of the Chartered College of Teachers, Impact 8.*

Conferences and Talks

2021: Poster accepted for Cog Sci 2021; poster accepted for SARMAC (The Society for Applied Research in Memory and Cognition); presentation of research EARLI SIG 16 Metacognition (European Association for Research into Learning and Instruction); talk to M.Phil. students, Cambridge; panel chair and contributor INSTRUCT symposium; BCCCD21 presentation (Budapest Central European University Conference on Cognitive Development).
2020: P.G.C.E. primer Cambridge 2020; presentation at Kaleidoscope, Cambridge; talk to M.Ed. students, Cambridge; FERSA (Faculty of Education Research Students Association)

2018: Poster Cog Sci; Poster EARLI; talk to Girls School Association, 2018; presentation to Headmasters and Headmistresses' Conference schools 2018.

2016 / 2017: Presentations and posters at the annual EdD conference, Cambridge

Other Publications / Peer Reviews

Barsham, H. (2021). Testing in Schools. *Academia Letters,* Article 624: https://doi.org/10.20935/AL624
Blog proposal accepted for Learning Scientists.org
2021 Peer reviewer for CERJ (Cambridge Educational Research e-journal)

Abstract

The testing effect is a robust cognitive strategy for improving recall from memory. Most of the literature about the testing effect (repeated retrieval practice) focusses on retrieval practice of academic, subject content. Test anxiety can occur in many individuals and from age seven upwards and students who are high in test anxiety account for 10-20% of a school population. The intervention in this study used retrieval practice to teach 10-to 11- year-olds about the testing effect itself; in effect, using the testing effect to teach the testing effect to try and increase feelings of self-efficacy in test taking and reduce the *worry* or cognitive element of test anxiety. The impact of the intervention was measured with a new Self-efficacy in Test Taking instrument and an adapted version of the *thoughts'* subscale from the Children's Test Anxiety Scale using pre-test and post-test participant responses. The intervention was delivered in the classroom over a period of six weeks and aimed to teach 10-to 11- year-olds to believe in their test taking abilities, that their "testing routes were well-oiled". There was a control group. The results show that for both the intervention and control group, the worry element of test anxiety reduced and self-efficacy in test taking increased at the post-test stage. The intervention increased feelings of test taking self-efficacy significantly for students who are high in test anxiety when measured with the Little Self-efficacy in Test Taking subscale.

ACKNOWLEDGEMENTS

I am deeply indebted to Dr. Michelle Ellefson, my supervisor for her expertise, guidance, rigour, time, care and patience. I would like to thank all the members of the INSTRUCT lab for the learning, sharing and supporting and to the Faculty of Education Quants group who are an invaluable, statistical resource group at Cambridge; to my EdD colleagues for their erudite emojis and comradery; to Tanya for her relentless belief in my 'crushing it'; to Claudia for answering my relentless questions about statistics.

My thanks too, to Prof. David Putwain, who never hesitated to send me his latest manuscript, which was so immensely helpful. To Doug Wren also for his help with the Children's Test Anxiety Scale measure. To Holmes Finch for getting back to me about MIMIC models. Thank you too to Jan Vermunt and Ros McLellan for their interest in my work and the feedback from my upgrade viva. Thank you to Ayesha Ahmed, my advisor. To Ricardo Sabates for my chat about the multilevel linear model.

Thank-you to my examiners for their commitment and time taken in reading this thesis.

Huge appreciation goes to the schools and especially the teachers and students who took part in the study. I could not have done it without you! I know how busy you are! Your data are essential to this study and very appreciated.

And finally, thank you to my family, especially my husband who can spot where I have left an apostrophe out on RStudio© and saved me hours of work; to my daughters for holding onto their conversations until I had 'just finished this bit'; to mum and dad for their constant and unconditional love; to my friends, who have not seen me for coffee, drinks and fun (even when we were allowed to go out) and understood why; to Pluto, my constant companion.

Finally, to the students and colleagues, who have listened to the mantras "study, test, test, test" and 'your testing routes are well-oiled' and tried it out

Table of Contents:

Abstract 3

List of Tables vi

List of Figures xvi

1. Introduction 2

1.1 Can Retrieval Practice of The Testing Effect Increase Self-efficacy in Tests and Reduce Test Anxiety, in 10-to 11- Year-olds? 2

1.2 Overview of the Thesis 5

2. Literature Review 9

2.1 The Desirable Difficulty of the Testing Effect 9
2.1.1 Retrieval Practice with 10-to 11-year-olds 11
2.1.2 Desirable Difficulties 20
2.1.3 Possible Solutions for Improving Students' Metacognition 24
2.1.4 The Testing Effect and Test Anxiety 27
2.1.5 Summary and Next Section 31

2.2 Test Anxiety 33
2.2.1 The Test Anxiety Construct 36
2.2.2 The Measures 42
2.2.3 Effects of Test Anxiety 44
2.2.4 Summary and Next Section 48

2.3. Self-efficacy 48
2.3.1 Theories and Models 49
2.3.2 Test Anxiety, Uncertain Control, Academic Buoyancy and Well-being 51
2.3.3 Creating a Self-efficacy Measure; the Self-efficacy in Test Taking Instrument 58
2.3.4 Summary and Next Section 59

3. Intervention 61

3.1 Test Anxiety and Interventions 62
3.1.1 Intervening in Cognitive Test Anxiety 64
3.1.2 Writing About Test Anxiety 65
3.1.3 Controlling Control for Self-efficacy in Test Taking and Well-being 66
3.1.4 The Design of My Intervention 67

3.2 Design-Based Theory - Intervention 68
3.2.1 Fidelity of Implementation 71
3.2.2. Effectiveness and Efficacy 79
3.2.3 The Delivery of My Intervention 79
3.2.4 A Mini Summary of the Intervention 80

4. Research Questions 82

4.1 The Research Questions 82

5. Pilot Study **89**

5.1 Purpose **89**

5.2 Method **89**
5.2.1 Participants 89
5.2.2 Pre and Post-test Measures 89
5.2.3 Materials 90
5.2.4 Procedure for Data Collection 91
5.2.5 The Pilot Intervention 91

5.3 Results **92**
5.3.1 Interviews 93
5.3.2 Discussion 94
5.3.3 Limitations 95
5.3.4 Lessons Learned from the Pilot 95

6. Method **98**

6.1 Method – Rationale **99**
6.1.1 Participants 99
6.1.2 Missing Data 100
6.1.3 Materials 100
6.1.4 Procedures 102

6.2 Design **103**

6.3 Ethical Considerations **105**

7. Data Analytic Plan and Preliminary Analyses **107**

7.1 Overview of Chapter 7 **107**
7.1.1 Evaluating the Self-report Questionnaires 114
7.1.2 Evaluating the Intervention 119

7.2 Part 1, The Children's Test Anxiety Scale Data **119**
7.2.1 Descriptive Statistics for the Individual Items 120
7.2.2 Factor Analysis 121
7.2.3 Measurement Invariance 136

7.3 Part II: Self-efficacy In Test Taking **139**
7.3.1 Descriptive Statistics for the Individual Items 139
7.3.2 Factor Analysis 140
7.3.3 Measurement Invariance for the Self-efficacy Test Taking Instrument (SETTI) 152

7.4 Intervention Data **153**
7.4.1 Assumptions 153

8. Results **157**

8.1 The Children's Test Anxiety Scale Results **157**
8.1.1 Latent Means Analysis 158
8.1.2 Multilevel Linear Analysis for the Children's Test Anxiety Scale 158

8.2 Self-efficacy in Test Taking Results **160**
8.2.1 Latent Means Analyses 161
8.2.2 Mixed 2 x 2 Factorial ANOVA on the Little Self-efficacy Scale in Test Taking Subscale 161

8.2.3 Latent Means Analyses 162
8.2.4 Multilevel Linear Analysis for the Self-efficacy in Test Taking Scale 163
8.2.5 ANCOVA for the Little Self-efficacy in Test Taking Scale 165

8.3 Research Question 2b 166
8.3.1 Multiple Indicators Multiple Causes – MIMIC Models 166

8.4. Students Who Are High in Test Anxiety 168
8.4.1 Research Question 3 168

8.5 Research Questions That Arose from This Study 171
8.5.1 Research Question 4 171
8.5.2 Research Question 5 172

9. Discussion 173

9.1 Worry Reduces in Both the Intervention and Control Groups 173
9.1.1 Interpretation of Results 173
9.1.2 Links to Test Anxiety 175
9.1.3 The Multilevel Linear Model 177

9.2 Self-efficacy in Test Taking Increased in the Intervention and the Control Group 179
9.2.1 Interpretation of Results 179
9.2.2 Links to Control 181
9.2.3 Limitations to Research Questions 1 and 2a 183
9.2.4 Implications for the Classroom from Research Questions 1 and 2a 186

9.3 MIMIC Models and the Relationship Between Test Anxiety and Self-efficacy in Test Taking. 188
9.3.1 Interpretation of results 188
9.3.2 Limitations 190
9.3.3 Implications 190

9.4 Students Who Are High in Test Anxiety Benefit Significantly from the Cognitive Intervention. 191
9.4.1 Interpretation of Results 191
9.4.2 Links to High Test Anxiety 193
9.4.3 Limitations 193
9.4.4 Implications for the Classroom 193

9.5 The Additional Research Questions. 194
9.5.1 Interpretation of results 195
9.5.2 Limitations 195
9.5.3 Implications 196

9.6 Conclusion 196

10 Reflections and Conclusions 198

10.1 Reflections About the Intervention 198
10.1.1 What Went Well 200
10.1.2 Even Better If 201
10.1.3 Next Steps 203

10.2 General Reflections About the Feedback, Results and Study 203

10.3 Recommendations – General, Policy, Reform? 205
10.3.1 Writing About Test Anxiety 206
10.3.2 Training About Fear Appeals 207
10.3.3 Proposal for Curricula Reform 207

10.4 Future Directions 209

10.5 Conclusion About the Intervention and Final Reflections 210

References 212

Appendix A – Questionnaires 245

Appendix B Teacher's Manual for the Intervention 247

Appendix C New Theory of Disuse 262

Appendix D Teacher and Student Feedback 265

Appendix E Ethics 274

Appendix F Results from the Pilot 282

Appendix G Results of the Children's Test Anxiety Scale 285

Appendix H Results of the Self-efficacy in Test Taking Instrument 342

Appendix I R Studio© Code 385

List of Tables

Table 3.1 *Meeting Design-Based Theory (Mckenney & Reeves, 2013)* 70

Table 3.2 *Fidelity of Implementation, Dusenbury et al., (2003)* 71

Table 3.3 *Detrich (1999) Contextual Variables* 73

Table 3.4 *Caroll et al. (2007) Conceptual Framework* 74

Table 3.5 *Field and Hole (2003) Threats to and Management of External Validity* 76

Table 3.6 *Field and Hole (2003) Threats to and Management of Internal Validity* 77

Table 3.7 *Bracht and Glass Ecological Validity (1968)* 78

Table 3.8 *Mini Summary of the Intervention* 80

Table 5.1 *Pilot - Validity of Questionnaire, N=33* 93

Table 6.1 *Schools, High Stakes Tests, Sample Number Used and Missing Data* 101

Table 7.1 *Descriptive Statistics for Children's Test Anxiety Data and Self-efficacy in Test Taking Data* 108

Table 7.2 *Analyses and Assumptions for Research Question 1: Does the Intervention that uses Retrieval Practice to Teach the Testing Effect Reduce Feelings of Worry in 10-to11-year-olds?* 109

Table 7.3 *Analyses and Assumptions for Research Question 2a: Does the Intervention Increase Feelings of Self-efficacy in Test Taking in 10-to-11-year-olds?* 111

Table 7.4 *Analyses and Assumptions for Research Question 2b: What do Self-efficacy in Test Taking Sum Scores Tell Us About Individual Test Anxiety?* 112

Table 7.5 *Analyses and Assumptions for Research Question 3: Do Feelings of Worry Predict How Well Students Self-efficacy Changes in the Intervention Group?* 112

Table 7.6 *Analyses and Assumptions for Research Question 4: Is There a Significant Difference Between the Groups in the Intervention Group Where the Researcher Delivered the Intervention?* 113

Table 7.7 *Analyses and Assumptions for Research Question 5: The SARS-CoV-2 Question. Is There a Difference in Feelings of Test Anxiety Between Students Who Sit High*

Stakes' Paper Tests or Online Tests? (*N = 51, n = 30 online tests, n = 21 paper tests*)..... 113

Table 7.8 *Skew and Kurtosis Values for Some Items on the Children's Test Anxiety Scale Data, n = 85* 120

Table 7.9 *Fit Indices Statistics for the Original Thoughts' Subscale with the Children's Test Anxiety Scale Data in This Study, n = 85, N = 340.*..... 122

Table 7.10 *Cronbach's Alpha for the Original Thoughts' Subscale, n = 85, N = 340*..... 123

Table 7.11 *Eigenvalues >1 for the Children's Test Anxiety Scale Data, n = 85, N = 340*... 128

Table 7.12 *Fit Indices Statistics for Exploratory Factor Analysis Using Lavaan© in RStudio©, n = 85.*..... 129

Table 7.13 *Factor Loadings for the Different Conditions - Children's Test Anxiety Scale for This Study* 130

Table 7.14 *Fit Indices Statistics for the Three Different Versions of the Adapted Thoughts' Subscale with the Children's Test Anxiety Scale Data (n = 85)* 134

Table 7.15 *Fit Indices Statistics for the Three Different Versions of the Adapted Thoughts' Subscale with the Children's Test Anxiety Scale Data (N = 340)*..... 135

Table 7.16 *Skew and Kurtosis Values of Items 5 and 6 for the Self-efficacy in Test Taking Data, n = 59 for the Intervention Group and n = 86 for the Control Group*..... 140

Table 7.17 *Cronbach's Alpha for the Self-efficacy in Test Taking Instrument - Intervention and Control Groups, n = 59 intervention, n = 86 control.* 142

Table 7.18 *Eigenvalues >1 for The Self-efficacy Data, n =59 (intervention), n = 86 (control), N = 290*..... 144

Table 7.19 *Exploratory Factors Analysis: Fit Statistics for the Self-efficacy Instrument Across Intervention and Control Groups and at Pre-test and Post-test Timepoints for One and Two Factor Extractions.* 146

Table 7.20 *Cronbach's Alpha for the Self-Efficacy in Test Taking Instrument* 148

Table 7.21 *Fit Indices for the Proposed Model for the Self-efficacy in Test-Taking Scale N=290*..... 149

Table 8.1 *Results of Multiple Indicators Multiple Causes(MIMIC) Models for the Adapted Thoughts' Subscale* 167

Table F.1 *Pilot - Paired t Test Results* 282

Table G.1 *Means, Standard Deviations, Skew and Kurtosis, Pre-test, Intervention Group (n = 85)* 285

Table G.2 *Means, Standard Deviations, Trimmed Mean, Skew and Kurtosis, Post-test, Intervention Group (n = 85)* 286

Table G.3 *Means, Standard Deviations, Trimmed Mean, Skew and Kurtosis, Pre-test, Control Group (n = 85)* 287

Table G.4 *Means, Standard Deviations, Skew and Kurtosis, Post-test, Control Group (n = 85)* 288

Table G.5 *Mean, Skew and Kurtosis for the Original Thoughts' Subscale, Pre-test, Intervention Group (n = 85)* 289

Table G.6 *Mean, Skew and Kurtosis for the Original Thoughts' Subscale, Post-test, Intervention Group (n = 85)* 290

Table G.7 *Mean, Skew and Kurtosis for the Original Thoughts' Subscale Pre-test, Control Group (n = 85)* 291

Table G.8 *Mean, Skew and Kurtosis for the Original Thoughts' Subscale, Post-test, Control Group (n =85)* 292

Table G.9 *Shapiro-Wilk Test for Normality of Distribution for Pre and Post Timepoints and Intervention and Control Group (N = 340)* 301

Table G.10 *Kaiser-Meyer-Olkin Measures of Sampling Adequacy (N = 340)* 301

Table G.11 *Bartlett's Test for Sphericity for the Children's Test Anxiety Scale instrument (N = 340)* 302

Table G.12 *Determinant Matrix Values for the Groups and Timepoints (N = 340)* 302

Table G.13 *Factor Loadings from the Original Children's Test Anxiety Scale Development Paper, Wren and Benson, (2004)* 307

Table G.14 *Condition 1: Exploratory Factor Analysis Showing the Factor Loadings for Three Factors with Pre-test Intervention Group (n = 85)* 308

Table G.15 *Condition 2: Exploratory Factor Analysis Showing the Factor Loadings for Four Factors With the Pre-test Intervention Group (n = 85)* 309

Table G.16 *Condition 3: Exploratory Factor Analysis Showing the Factor Loadings for Three Factors with Post-test Intervention Group (n = 85)* 310

Table G.17 *Condition 4: Exploratory Factor Analysis Showing the Factor Loadings for Four Factors Post-test Intervention Group (n = 85)* 311

Table G.18 *Condition 5: Exploratory Factor Analysis Showing the Factor Loadings for Three Factors Pre-test Control Group (n = 85)* 312

Table G.19 *Condition 6: Exploratory Factor Analysis Showing the Factor Loadings for Four Factors Pre-test Control Group (n=85)* 313

Table G.20 *Condition 7: Exploratory Factor Analysis Showing the Factor Loadings for Three Factors Post-test Control Group (n = 85)* 314

Table G.21 *Condition 8: Exploratory Factor Analysis Showing the Factor Loadings for Four Factors Post-test Control (n = 85)* 315

Table G.22 *Three and Four Factor Loadings and Cronbach's Alpha for the Original Thoughts' Subscale Across All Conditions (N = 340)* 316

Table G.23 *Measurement Invariance of Adapted Thoughts' Subscale over Timepoints Pre-test and Post-test (N = 340)* 317

Table G.24 *Measurement Invariance of Adapted Thoughts' Subscale for Intervention and Control Groups (N = 340)* 318

Table G.25 *Measurement Invariance of Adapted Thoughts' Subscale Across Gender Groups (N = 340)* 319

Table G.26 *ANOVAs for Timepoint Measurement Invariance (N = 340)* 320

Table G.27 *ANOVAs for Intervention and Control Group Measurement Invariance (N = 340)* 321

Table G.28 *ANOVAs for Gender Group Measurement Invariance (N = 340)* 322

Table G.29 *Levene's Test for Homogeneity of Variance for the Adapted Thoughts' Subscale* 323

Table G.30 *Results for Within Subjects Effects Timepoint for the Adapted Thoughts' Subscale (N = 340)* 323

Table G.31 *Results for Between Subjects Effects Group Intervention and Control (N =340)* 324

Table G.32 *Group Descriptive Statistics for the ANOVA (RQ1) (N = 340)* 324

Table G.33 *Latent Means' Fit Statistics of Adapted Thoughts' Subscale Across the Timepoints Pre and Post for the pre-test and post-test Intervention Groups (n = 170)* 325

Table G.34 *Latent Means' ANOVA for Adapted Thoughts' Subscale Across the Timepoints Pre and Post for the pre-test and post-test Intervention Groups (n = 170)* 325

Table G.35 *Latent Means' Fit Statistics for Adapted Thoughts' Subscale Across the Timepoints Pre and Post for the pre-test and post-test Control Groups (n = 170)* 326

Table G.36 *Latent Means' ANOVA for the Adapted Thoughts' Subscale Across the Timepoints Pre and Post for the pre-test and post-test Control Groups (n = 170)* 326

Table G.37 *Models, Equations and Key Comparative Statistics for Multilevel Linear Model Analysis* 327

Table G.38 *Step 1 Information Criteria* 328

Table G.39 *Step 1 Fixed Effects - Type III Tests of Fixed Effects* 328

Table G.40 *Step 1 Estimates of Fixed Effects* 328

Table G.41 *Step 1 Estimates of Covariance Parameters* 329

Table G.42 Step 2a Information Criteria 329

Table G.43 *Step 2a Fixed Effects - Type III Tests of Fixed Effects* 329

Table G.44 *Step 2a Estimates of Fixed Effects* 330

Table G.45 *Step 2a Estimates of Covariance Parameters* 330

Table G.46 *Step 2b Information Criteria* 330

Table G.47 *Step 2b Fixed Effects - Type III Tests of Fixed Effects* 331

Table G.48 *Step 2b Estimates of Fixed Effects* 331

Table G.49 *Step 2b Estimates of Covariance Parameters* 331

Table G.50 *Levene's Test for Homogeneity of Variance for the Experimenter Effect / Fidelity Treatment; researcher groups vs teacher groups* 332

Table G.51 *Results for Within Subjects Effects for Timepoint for the Researcher Vs Teacher Groups Using the Adapted Thoughts' Subscale to Measure Test Anxiety (N = 62)* 332

Table G.52 *Results for Between Subjects Effects for Timepoint for the Researcher Vs Teacher Groups Using the Adapted Thoughts' Subscale to Measure Test Anxiety (N = 62)* 333

Table G.53 *Descriptive Statistics for Research Question 4 (N = 62)* 333

Table G.54 *Levene's Test for Homogeneity of Variance for Paper Vs Online High Stakes' Tests* 334

Table G.55 *Independent Samples T Test – Test Anxiety with High Stakes' Paper Tests vs Online Tests (N = 51)* 334

Table G.56 *Descriptive Statistics for Independent T Test – Test Anxiety with High Stakes' Tests vs Online Tests (N =51)* 334

Table G.57 *Multi Indicators Multiple Causes (MIMIC Models)* 335

Table H.1 *Means, Standard Deviations, Skew and Kurtosis for the Self-efficacy in Test Taking Instrument with the Pre-test, Intervention Group (n = 59)* 342

Table H.2 *Means, Standard Deviations, Skew and Kurtosis for Self-efficacy in Test Taking Instrument with the Post-test, Intervention Group (n = 59)* 343

Table H.3 *Means, Standard Deviations, Skew and Kurtosis for Self-efficacy in Test Taking Instrument with the Pre-test, Control Group (n = 86)* 344

Table H.4 *Means, Standard Deviations, Skew and Kurtosis for Self-efficacy in Test Taking Instrument with the Post-test, Control Group (n = 86)* 345

Table H.5 *Shapiro-Wilk Test for Normality of Distribution for Pre and Post Timepoints and Intervention and Control Group with the Self-efficacy Data* (*N* = 290) 351

Table H.6 *Kaiser-Meyer-Olkin Measures of Sampling Adequacy for the Self-efficacy Instrument* (*N* = 290) 351

Table H.7 *Bartlett's Test for Sphericity for the Self-efficacy Instrument* (*N* = 290) 352

Table H.8 *Determinant Matrix Values for the Self-efficacy in Test Taking at Pre-test and Post-test with the Intervention and Control Groups.* (*N* = 290) 352

Table H.9 *Shapiro-Wilk Test for Normality of Distribution for Pre and Post Timepoints and Intervention and Control Group for the Self-efficacy in Test Taking Instrument* (*N* = 290) 353

Table H.10 *Factor loadings of items onto one and two factor models for groups and timepoints for the self-efficacy data. (N = 290)* 362

Table H.11 *Measurement Invariance for the Self-efficacy Test-Taking Subscale SETT Across Timepoint Pre-test and Post-test.* (*N = 290)* 363

Table H.12 *ANOVA Between the Configural and Fully Constrained Models for the SETT subscale Across Timepoint, Pre-test and Post-test.* (*N = 290)* 363

Table H.13 *Measurement Invariance for the Self-efficacy Test-Taking Subscale SETT Across Intervention and Control Groups.* (*N = 290)* 364

Table H.14 *ANOVA Between the Configural and Constrained Models for the SETT subscale Across Intervention and Control Groups.* (*N = 290)* 365

Table H.15 *Measurement Invariance for the Self-efficacy Test-Taking Subscale SETT Across Gender Groups. (N = 290)* 365

Table H.16 *ANOVA Between the Configural and Constrained for the SETT Subscale Across Gender Groups. (N = 290)* 366

Table H.17 *Measurement Invariance for the Little Self-efficacy in Test-Taking Subscale Across Timepoint, Pre-test and Post-test. (N = 290)* 366

Table H.18 *ANOVA Between the Configural and Fully Constrained Models for the LSETT Subscale Across Timepoint Pre-test and Post-test. (N = 290)* 367

Table H.19 *Measurement Invariance for the Little Self-efficacy Test-Taking Subscale Across Intervention and Control Groups. (N = 290)* 368

Table H.20 *ANOVA Between the Configural and Fully Constrained Models for the LSETT subscale Across Intervention and Control Groups. (N = 290)* 368

Table H.21 *ANOVA Between the Constrained and Metric Models for the LSETT subscale Across Intervention and Control Groups. (N = 290)* 369

Table H.22 *Measurement Invariance for the Little Self-efficacy Test-Taking Subscale Across Gender Groups.* (*N = 290)* 369

Table H.23 *ANOVA Between the Configural and Fully Constrained Models for the LSETT subscale Across Gender Groups. (N = 290)* 370

Table H.24 *Levene's Test for Homogeneity of Variance - Self-efficacy in Test Taking*..... 370

Table H.25 *Results for Within Subjects Effects for Timepoint for the Self-efficacy in Test Taking Instrument (N = 290)*..... 371

Table H.26 *Results for Between Subjects Effects for Group for the Self-efficacy in Test Taking Instrument (N = 290)*..... 371

Table H.27 *Descriptive Statistics for the Groups for the Self-Efficacy in Test Taking Instrument (N =290)*..... 371

Table H.28 *Latent Means Fit Statistics of Self-efficacy in Test Taking Subscale Across the Timepoints Pre and Post for Intervention (n = 118)*..... 372

Table H.29 *Latent Means'ANOVA for Self-efficacy in Test Taking Across the Timepoints Pre and Post Intervention Groups (n=118)* 372

Table H.30 *Latent Means' Fit Statistics of Self-efficacy in Test Taking Subscale Across the Timepoints Pre and Post Control Groups (n =172).*..... 373

Table H.31 *Latent Means' ANOVA for the Self-efficacy in Test Taking Subscale Across the Timepoints Pre and Post for Control Groups (n=172)* 373

Table H.32 *Levene's Test for Homogeneity of Variance - Little Self-efficacy in Test Taking* 374

Table H.33 *Results for Within Subjects Effects for Timepoint for the Little Self-efficacy in Test Taking Instrument (N = 290)* 374

Table H.34 *Results for Between Subjects Effects for Group for the Little Self-efficacy in Test Taking Instrument (N = 290)*..... 374

Table H.35 *Descriptive Statistics for the Groups for the Little Self-Efficacy in Test Taking Instrument (N = 290)* 375

Table H.36 *Post Hoc Test for Little Self-efficacy Subscale (N = 290)* 375

Table H.37 *Little Self-efficacy in Test Taking Levene's Test – ANCOVA, N = 145* 376

Table H.38 *Little Self-efficacy in Test Taking Subscale ANCOVA, N = 145* 376

Table H.39 *Little Self-efficacy in Test Taking Subscale Marginal Means* 376

Table H.40 *Latent Means' Fit Statistics for Little Self-efficacy in Test Taking Subscale Across the Timepoints Pre and Post for the Intervention Group (n =118)* 377

Table H.41 *Latent Means' ANOVA for the Little Self-efficacy in Test Taking Subscale Across the Timepoints Pre and Post for the Intervention Group (n =118)* 377

Table H.42 *Latent Means' Fit Statistics of Little Self-efficacy in Test Taking' Subscale Across the Timepoints Pre and Post Control Group (N =172)* 378

Table H.43 *Latent Means' ANOVA for the Little Self-efficacy in Test Taking Subscale Across the Timepoints Pre and Post for the Control Group (N =172)* 378

Table H.44 *Levene's Test for Homogeneity of Variance for High Test Anxiety Students Measured with the LSETT (Little Self-efficacy in Test Taking) Subscale* 379

Table H.45 *Results for Within Subjects Effects for Timepoint for the Little Self-efficacy in Test Taking Subscale with Students Who Are High in Test Anxiety (N = 31)* 379

Table H.46 *Results for Between Subjects Effects for Group for the Little Self-efficacy in Test Taking Subscale with Students Who Are High in Test Anxiety (N = 31)* 380

Table H.47 *Descriptive Statistics for the Groups for the Little Self-Efficacy in Test Taking Instrument with Students Who Are High in Test Anxiety (N = 31)* 380

Table H.48 *Post hoc Test for the Interaction Between Group and Timepoint – Little Self-efficacy Subscale (N =31)* 381

Table H.49 *Levene's Test for Homogeneity of Variance for High Test Anxiety Students Measured with the SETT (Self-efficacy in Test Taking) Subscale* 381

Table H.50 *Results for Within Subjects Effects for Timepoint for the Self-efficacy in Test Taking Subscale with Students Who Are High in Test Anxiety (N = 31)* 382

Table H.51 *Results for Between Subjects Effects for Group for the Self-efficacy in Test Taking Subscale with Students Who Are High in Test Anxiety (N = 31)*..... 382

Table H.52 *Descriptive Statistics for the Groups for the Self-Efficacy in Test Taking Instrument with Students Who Are High in Test Anxiety (N = 31)* 382

Table H.53 *Levene's Test for Homogeneity of Variance for High Test Anxiety Students Measured with the Adapted Thoughts' Subscale*..... 383

Table H.54 *Results for Within Subjects Effects for Timepoint for the Adapted Thoughts' Subscale with Students Who Are High in Test Anxiety (N = 31)*..... 383

Table H.55 *Results for Between Subjects Effects for Group for the Adapted Thoughts' Subscale with Students Who Are High in Test Anxiety (N = 31)*..... 384

Table H.56 *Descriptive Statistics for the Groups for Adapted Thoughts' Subscale with Students Who Are High in Test Anxiety (N = 31)*..... 384

List of Figures

Figure 1.1 *The Thinking and Design Behind this Experiment* 7

Figure 2.1 *The Mechanism* 41

Figure 2.2 *The S-REF Model - Self-referent Executive Function (Zeidner & Matthews, 2005)* 41

Figure 2.3 *Bandura (1986b) Triadic Reciprocal Causation* 49

Figure 4.1 *The S-REF Model - Self-referent Executive Function (Zeidner & Matthews, 2005)* 83

Figure 4.2 *MIMIC (Multiple Indicators, Multiple Causes) Model of Self-efficacy Covariate in Relation to the Adapted Thoughts' Subscale* 86

Figure 7.1 *The Adapted Thoughts' Subscale for the Children's Test Anxiety Scale Data, N =340* 138

Figure 7.2 *Self-efficacy in Test Taking Scale* 151

Figure 7.3 *Little Self-efficacy in Test Taking Scale* 151

Figure 8.1 *Means of Pre-test and Post-test Intervention and Control Groups for the Children's Test Anxiety Scale Data Measured Using the Adapted Thoughts' Subscale* 158

Figure 8.2 *Pre-test as a Covariate to the Post-test Scores for the Intervention and Control Groups Showing Non-homogeneity of Slopes and Clustering* 159

Figure 8.3 *Pre-test as a Covariate to the Post-test Scores for the Different Schools* 159

Figure 8.4 *Means of Pre-test and Post-test Intervention and Control Groups for the Self-efficacy in Test Taking Data Measured Using the Self-efficacy in Test Taking Subscale* 161

Figure 8.5 *Means of Pre-test and Post-test Intervention and Control Groups for the Little Self-efficacy Scale in Test Taking Data Measured Using the Little Self-efficacy in Test Taking Subscale* 162

Figure 8.6 *Pre-test as a Covariate to the Post-test Scores for the Intervention and Control Groups Showing Non-homogeneity of Slopes* 164

Figure 8.7 *Pre-test as a Covariate to the Post-test Scores for the Different Schools* 164

Figure 8.8 *Little Self-efficacy in Test Taking Subscale Groups* 165

Figure 8.9 *Little Self-efficacy in Test Taking Schools* 165

Figure 8.10 *The MIMIC (Multiple Indicators Multiple Causes) Model (Diagram from Finch and French, 2015) for the Adapted Thoughts Latent Variable with Self-efficacy in Test Taking Subscales as Exogenous Variables (Covariate)* 166

Figure 8.11 *Means of Pre-test and Post-test Students who are High in Test Anxiety for the Intervention and Control Groups Data Measured Using the Little Self-efficacy Scale in Test Taking Subscale* 169

Figure 8.12 *Means of Pre-test and Post-test Students who are High in Test Anxiety for the Intervention Groups Data Measured Using the Self-efficacy in Test Taking Subscale* 170

Figure 8.13 *Worry Measured by the Adapted Thoughts' Subscale for Students Identified as High in Test Anxiety* 171

Figure G.1 *Responses for the Original Thoughts' Subscale Pre-test Intervention Group, n =85* 293

Figure G.2 *Responses for the Original Thoughts' Subscale Post-test Intervention Group, n = 85* 293

Figure G.3 *Responses for the Original Thoughts' Subscale Pre-test Control Group, n = 85* 294

Figure G.4 *Responses for the Original Thoughts' Subscale, Post-test, Control Group, n =85* 294

Figure G.5 *Violin Plot - Responses for Adapted Thoughts' Subscale for Intervention Pre-test, n = 85* 295

Figure G.6 *Violin Plot – Responses for Adapted Thoughts' Subscale for Intervention Post - test, n = 85* 295

Figure G.7 *Violin Plot – Responses for Adapted Thoughts' Subscale for Control Pre-test, n = 85* 296

Figure G.8 *Violin Plot Responses for Adapted Thoughts' Subscale for Control Post-test, n = 85* 296

Figure G.9 *Children's Test Anxiety Scale Pre-test Intervention Group Correlations, n = 85* 297

Figure G.10 *Children's Test Anxiety Scale Post-test Intervention Group Correlations, n = 85* 298

Figure G.11 *Children's Test Anxiety Pre-test Control Group Correlations, n = 85* 299

Figure G.12 *Children's Test Anxiety Post-test Control Group Correlations, n = 85* 300

Figure G.13 *Children's Test Anxiety Scale Intervention Pre-test Correlations, (*X = *Significant Values)* 303

Figure G.14 *Children's Test Anxiety Scale Intervention Post-test Correlations (* X = *Significant Values)* 304

Figure G.15 *Children's Test Anxiety Scale Control Pre-test Correlations (*X = *Significant Values)* 305

Figure G.16 *Children's Test Anxiety Scale Control Post-test (X = Significant Values)* 306

Figure G.17 *Intervention, Pre-test with Self-efficacy in Test Taking Sum Score as Covariate* 336

Figure G.18 *Intervention Pre-test with Little Self-efficacy in Test Taking Sum Score as Covariate* 336

Figure G.19 *Control, Pre-test with Self-efficacy in Test Taking Sum Score as Covariate*.... 337

Figure G.20 *Control, Pre-test with Little Self-efficacy in Test Taking Sum Score as Covariate* 337

Figure G.21 *Intervention post-test with Self-efficacy in Test Taking Post-test Sum Score as Covariate* 338

Figure G.22 *Intervention post-test, Little Self-efficacy in Test Taking Post-test Sum Score as Covariate* 338

Figure G.23 *Control, Post-test with Self-efficacy in Test Taking Sum Score as Covariate* .. 339

Figure G.24 *Control, Post-test with Little Self-efficacy in Test Taking Sum Score as Covariate* 339

Figure G.25 *Intervention post-test with Self-efficacy in Test Taking Pre-test Sum Score as Covariate* 340

Figure G.26 *Intervention post-test with Little Self-efficacy in Test Taking Pre-test Sum Score as Covariate* 340

Figure G.27 *Control post-test with Self-efficacy in Test Taking Pre-test Sum Score as Covariate* 341

Figure G.28 *Control post-test with Little Self-efficacy in Test Taking Pre-test Sum Score as Covariate* 341

Figure H.1 *Individual Histograms of Each Item on the Self-efficacy in Test Taking Questionnaire -Pre* 346

Figure H.2 *Individual Histograms of Each Item on the Self-efficacy in Test Taking Questionnaire -Post* 346

Figure H.3 *Self-efficacy in Test Taking Intervention Group Pre-test, n = 59* 347

Figure H.4 *Self-efficacy in Test Taking Intervention Group Post-test, n = 59* 347

Figure H.5 *Self-efficacy in Test Taking Control Group Pre-test, n = 86* 348

Figure H.6 *Self-efficacy in Test Taking Control Group Post-test, n = 86* 348

Figure H.7 *Little Self-efficacy in Test Taking Intervention Group Pre-test, n = 59* 349

Figure H.8 *Little Self-efficacy in Test Taking Intervention Group Post-test, n = 59* 349

Figure H.9 *Little Self-efficacy in Test Taking Control Group Pre-test, n = 86* 350

Figure H.10 *Little Self-efficacy in Test Taking Control Group Post-test, n = 86* 350

Figure H.11 *Self-efficacy in Test Taking Intervention Pre-test, n = 59* 354

Figure H.12 *Self-efficacy in Test Taking Intervention Post-test, n = 59* 355

Figure H.13 *Self-efficacy in Test Taking Control Pre-test, n = 86* 356

Figure H.14 *Self-efficacy in Test Taking Control Post-test, n = 86* 357

Figure H.15 *Correlations for the Self-efficacy in Test Taking Instrument with the Pre-test Intervention Group, n = 59* 358

Figure H.16 *Correlations for the Self-efficacy in Test Taking Instrument with the Post-test Intervention Group, n = 59* 359

Figure H.17 *Correlations for the Self-efficacy in Test Taking Instrument with the Pre-test Control Group, n = 86* 360

Figure H.18 *Correlations for the Self-efficacy Instrument with the Post-test Control Group, n = 86* 361

1. Introduction

1.1 Can Retrieval Practice of The Testing Effect Increase Self-efficacy in Tests and Reduce Test Anxiety, in 10-to 11- Year-olds?

Taking tests is an integral part of life. An individual's ability is often measured through tests. The peak of test taking occurs during most school and university years. The problem is that taking tests can be a terrifying and debilitating experience for some individuals and can lead to poor performance in these situations.

As a teacher and headteacher (principal), I have witnessed the anxiety that can be generated by having to take high stakes' tests in students and parents. By high stakes' tests, I mean tests that have an important outcome. In this study, the high stakes' tests are taken by Year 6 to determine entrance to a chosen senior school and possible scholarship awards. In addition, some students took Key Stage 2, SATs (Statutory Assessment Tests), the results of which are often used to predict G.C.S.E. (General Certificate of Secondary Education) grades, five years later in Year 11 (15-to16-year-olds).

As a headteacher, in the independent (private) sector of education, I have sent messages home and delivered talks to parents about the need for a calm approach at these evaluative pinch-points of the year. In my capacity as Head of a Prep school (private school for 3- to 11-year-olds in England), I explained to parents, ahead of high stakes' tests, that if they, as parents, become tense and stressed about important evaluative situations, then it may increase anxiety about taking the test for their children.

I have seen how anxiety about upcoming high stakes' tests can go viral, spreading rapidly amongst peers (and some parents) using various social media. I wanted to change students' perception about taking tests so that they felt more confident in their test taking abilities in evaluative situations and encourage positive comments and feelings rather than negative ones. Taking tests seems to be a feature of everyday life, especially while doing formal education or training.

I started this research project because I wanted to improve children's experience of taking tests. I spent my holidays reading about the brain and how learning happens. I became very interested in how memory works and the impact of retrieval practice. I did not know at the time of starting my research how important the subject of retrieval practice, *the testing effect* was to become nationally to educators. I did know that there was very little knowledge about retrieval practice with primary school students.

The Initial Teacher Training core framework in the U.K. has just included retrieval practice in one of its teaching standards. From September 2020, trainee teachers will be expected to use retrieval practice in their classroom to qualify as teachers. I welcome the addition as I have long been aware of the benefits of retrieval practice for children (and teachers). This study is about retrieval practice with 10-to 11- year-olds in relation to test anxiety and is breaking new ground on many levels with its focus on this age-group and use of retrieval practice working in a hybrid way as a pastoral and academic intervention in test anxiety. This is an important study, as the benefits of retrieval practice have now been recognized nationally with its inclusion into the Initial Teacher Training core framework, but the evidence to draw upon is limited. Thus, this study can contribute towards the evidence base that is needed for primary school teachers.

Furthermore, retrieval practice is used in the intervention of this study to address the problem of test anxiety to give students a coping strategy when taking tests and is not focused on the academic subject content recall that is the focus of most retrieval practice studies, although practice tests were happening simultaneously.

Currently, in England, we have high stakes SATs (Statutory Assessment Tests) in state primary schools for Year 2 (6- to 7-year-olds) and Year 6. For Year 6 students, (10-to11-year-olds), another type of high stakes' test is the 'eleven plus' exam. The 'eleven plus' exams determine acceptance into the U.K. grammar schools' system. This is a state system that operates in some areas of the U.K. and entrance to these highly sought after schools is determined purely from the scores of the 'eleven plus', exam.

In the independent (private) sector of English education, there are many scholarships available for various students, including Year 6. Most of these scholarships rely on scores from independently set, independent (private) school exams and can cover up to 50% of school fees (tuition). They can also determine which independent, senior school (also called secondary school, 11-to 18-year-olds) one is accepted at. All these high stakes' tests can be nerve-wracking experiences for some students.

I am motivated to make a positive difference for students by trying to increase feelings of self-efficacy in taking tests and thus reducing the worry or cognitive aspect of test anxiety. Given that testing is a regular feature of school life, it is important to prepare students specifically for these evaluative situations. Preparation for tests is usually done through taking practice tests. However, knowing why practice testing works seems to be the

missing piece of the testing puzzle for classroom teaching. From my experience, it is important for students to understand how and why testing works.

I developed this project because it is my expectation that if students learn about the testing effect, and are able to remember its benefits, then they will feel more confident and self-efficacious when taking tests. My goal was to create an intervention that is easily administered by a classroom teacher in a short, weekly session over a period of six weeks to teach the idea about the testing effect to Year 6 students (10-to 11-year-olds). I selected this age-group because of the high stakes' tests that occur during Year 6.

As an experienced teacher, I know that relating student learning to context is always essential. If a student can understand 'why' they are learning a particular 'thing' and can understand the relevance to them, then, usually, they are more motivated to learn it. (e.g., Zepeda et al., 2020). When students understand how their knowledge of a particular strategy, in this case, how the testing effect works, they can link the strategy to the evaluative situation and the educator is in a position of having provided high control (Putwain & Aveyard, 2018) so that, in this case, the students can draw upon positive self-beliefs.

In essence, the design of this thesis draws upon of three different theoretical fields: the cognitive theory of the testing effect, (e.g., Bjork, 1975; Roediger & Karpicke, 2006b), self-efficacy theory (e.g., Bandura, 1997) and test anxiety theory (e.g., Zeidner,1998). The literature from these three fields is more extensive than I have space within this thesis. As such, I have focused my literature review on the key aspects of these three areas that I found essential for this study, rather than a full overview of each field.

My study is based on the principles of a design experiment and compares students who completed an intervention that uses retrieval practice to teach the testing effect to other students in a control group that continued their normal classroom activities. I compared these groups on two core aspects: feelings of worry or cognitive test anxiety and feelings of test taking self-efficacy. To measure feelings of test anxiety, I used an adapted version of the Children's Test Anxiety Scale, *thoughts*' subscale (Wren & Benson, 2004), to focus on the concept of *worry* because worry is viewed as the cognitive element of test anxiety. To measure feelings of test taking self-efficacy, I developed a new questionnaire because none were available for this age-group.

To my knowledge, this type of intervention and these measures have not been researched before and with this age-group. Taking this research approach produced four main research questions:

Research Question 1: Does the intervention that uses retrieval practice to teach the testing effect reduce feelings of worry in 10-to11-year-olds?

Research Question 2a: Does the intervention increase feelings of self-efficacy in test taking in 10-to-11-year-olds?

Research Question 2b: What do self-efficacy in test taking sum scores tell us about individual test anxiety?

Research Question 3: Do feelings of worry predict how well student self-efficacy changes in the intervention group?

A further two questions were added as a result of thinking and life circumstances that occurred during the EdD process. Firstly, I questioned the fidelity of the experiment in my role as a researcher and potential bias and as a result, I conducted a comparative analysis about a reduction of feelings of the worry element of test anxiety between the groups, where I administered the intervention, and other, similar groups, where the intervention was administered by teachers. The has resulted in Research Question 4. Secondly, as some of the data in the control group had to be collected using online high stakes' tests, as a result of the pandemic, I include a comparison of the worry element of test anxiety in relation to whether the students took high stakes' tests online or the more traditional paper tests.

Research Question 4: Is there a significant difference between the groups in the intervention group where the researcher delivered the intervention?

Research Question 5: The Covid SARS COV 2 Question. Is There a Difference in Feelings of Test Anxiety Between Students Who Sit High Stakes' Paper Tests or Online Tests?

A small amount of feedback was collected in the form of teacher comments about the intervention and teacher notes from pupil conversations after the post-test questionnaires. (see Appendix D)

1.2 Overview of the Thesis

This thesis explores using retrieval practice of the testing effect in an intervention to increase self-efficacy in test taking and reduce the cognitive or worry element of test anxiety.

Other desirable difficulties (Bjork & Bjork, 2011) such as spacing and interleaving are introduced in the intervention and may be a form of differentiation by allowing different learners to access further knowledge about how learning can work. (see comments in Chapter 10) However, the main message to the 10-to 11-year-olds in the intervention is about the power of "study, test, test, test" (Roediger & Karpicke, 2006b) and as such, the literature review begins with the testing effect.

Problems with students monitoring their learning effectively and the importance of being able to monitor studies well are ideas that are inextricably linked to the creation of a successful intervention in this study. Therefore, judgements of learning (Bjork et al., 2013) effortful learning and error in learning are considered, to give the reader the insights into the importance of teaching students about the testing effect. In this study, the saying 'Your testing routes are well-oiled' is the catchphrase used in the intervention and encapsulates the idea of retrieval practice being a memory modifier (Bjork, 1975).

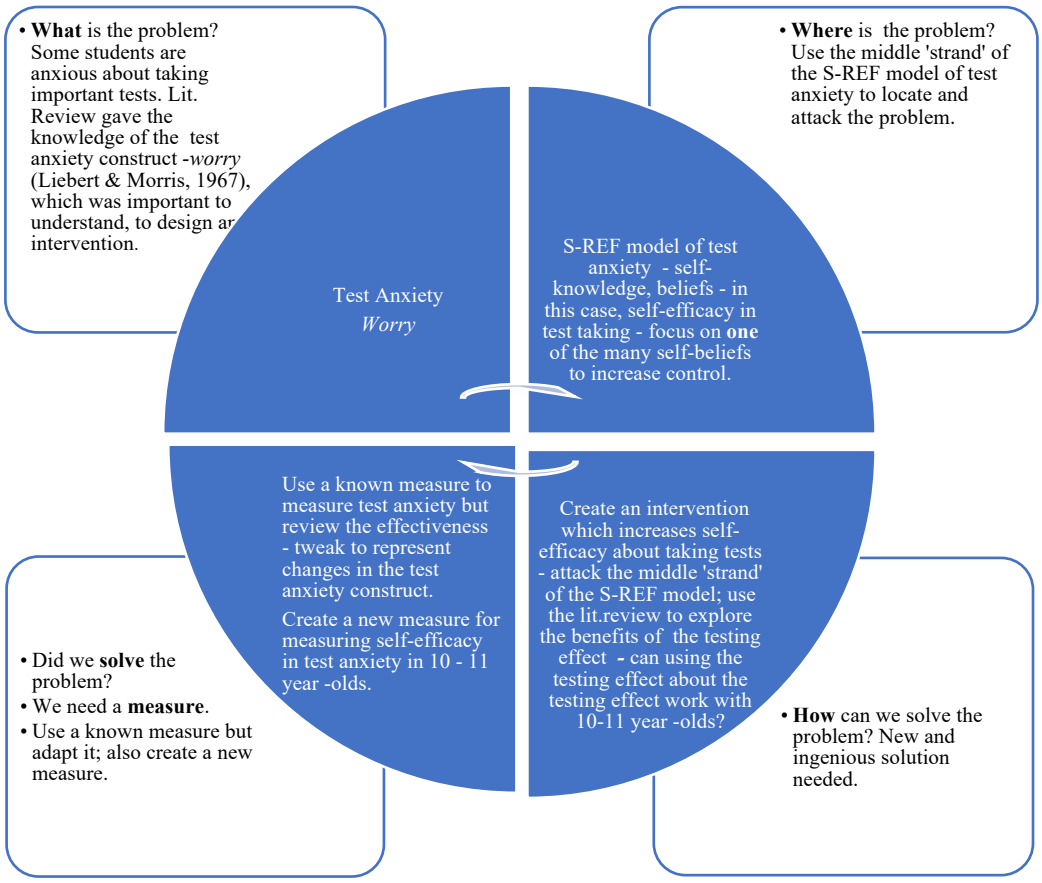
Test anxiety theory and self-efficacy theory are discussed as they are the other two key theories that are foundational to this study. Test anxiety theory is explored to find out about the test anxiety construct. It was important for me to understand how test anxiety works so that I could design an optimal intervention using the testing effect. I review some of the different models and explore the most relevant measure to this study. Interventions for test anxiety are reviewed. This study focusses on the worry or cognitive element of test anxiety (Liebert & Morris, 1967).

Self-efficacy sits under the umbrella of social cognitive theory (Bandura, 1997) and is explored in context to the intervention and the creation of a new instrument, Self-efficacy in Test Taking. Self-efficacy can be found in the self-knowledge beliefs strand of the S-REF (Self-referent executive function) model of test anxiety (Zeidner & Matthews, 2005).

The intervention in this study has been designed to try and increase cognitive control using self-efficacy in test taking beliefs to manage self-knowledge beliefs more effectively to try and reduce the worry, or cognitive element of test anxiety.

Figure 1.1

The Thinking and Design Behind this Experiment



Following the literature review, there are four main research questions that answer the question 'Can Retrieval Practice of the Testing Effect Increase Feelings of Self-efficacy and Reduce Test Anxiety in 10-to11-year-olds?' These are followed by Research Question 4 Research Question 5. The pilot study is included, and lessons learned from this trial shape the main study. The method chapter for the main study has an increased sample size for measuring a reduction in test anxiety ($N = 170$) and includes a control group. The measure is different in that only the *thoughts'* subscale, in the form of an *Adapted Thoughts'* subscale from the Children's Test Anxiety Scale to measure the worry component of test anxiety is used. I also created a new instrument to measure Self-efficacy in Test Taking ($N = 145$)

Preliminary Analyses (I, II) detail my decision making in reducing the data for the main analysis. Preliminary Analysis I is about the Children's Test Anxiety Scale and the decision I made to break with convention by exploring and confirming factors in a known measure. Preliminary Analysis II details the exploratory and confirmatory factor analyses to

identify the subscales for the new Self-efficacy in Test Taking Instrument. Preliminary Analyses III examines any further assumptions prior to answering the research questions. Throughout the preliminary analyses, I try to explain the rationale behind decisions that have been taken. The Results chapter provides the reader with the answers to the research questions. The thesis ends with an interpretation of the results, details the limitations of the study and reflects upon the implications for schools. Possible future research and school reforms are also considered.

This study is innovative and makes some original contributions to the fields of the cognitive theory of the testing effect, test anxiety and self-efficacy. An *Adapted Thoughts'* subscale was created to measure the worry element of test anxiety for this data. An innovative intervention used retrieval practice of the testing effect to try to create feelings of self-efficacy in testing in 10-to11-year-olds. A new measure for Self-efficacy in Test Taking was created to measure the impact of the intervention on feelings of self-efficacy in taking tests.

This study is the result of combining my roles as headteacher and researcher to try and improve outcomes for students. My voice as a teacher and headteacher (principal) is evident throughout this thesis, as this work is the result of a passionate educator who wants the very best for students and has seen, first-hand the anxiety that some students feel, about taking tests.

2. Literature Review

In this literature review, I consider three theoretical fields that are the foundations of this study: the testing effect, (e.g., Roediger & Karpicke, 2006b); test anxiety (e.g., Zeidner, 1998) and self-efficacy (Bandura, 1997). I have experimented with the idea that these three theories can work reciprocally to reduce cognitive test anxiety in 10-to11- year-olds.

Most of the literature about the testing effect and test anxiety is from research with older students and I have had to draw on this research extensively in the absence of age relevant research. I will highlight where there is age relevant research to use.

I am particularly interested in the idea of *perceived control*, the degree of certainty a student has about how to achieve well and avoid bad results (Putwain & Aveyard, 2018) and *uncertain control,* one of the self-beliefs that contributes towards perceiving high stakes' tests as threatening (Putwain & Pescod, 2018) for students, in evaluative situations. Perceived control means that a student understands that their action and efforts can affect academic outcomes (Putwain & Aveyard, 2018) and uncertain control means that a student cannot relate their actions and efforts to academic outcomes (Putwain & Pescod, 2018). Putwain (2018) suggests that perceived control and uncertain control form part of the self-beliefs' knowledge strand of the S-REF (Self-referent executive function) model of test anxiety (Zeidner & Matthews, 2005).

My aim is that the intervention in this study that uses the testing effect to teach the testing effect will reduce the worry element of test anxiety by increasing feelings of self-efficacy in test taking in 10-to 11-year-olds and that in doing so, I provide some much needed research for this age-group into test anxiety and retrieval practice. (Barsham & Ellefson, 2020) It is important to note at the outset of this thesis that there is a "pressing need" for an intervention for test anxiety with this age-group. (Putwain, 2020, p.6)

2.1 The Desirable Difficulty of the Testing Effect

The testing effect or repeated retrieval practice is a robust strategy for enhancing later retention of information to be recalled from long term memory (e.g., Bjork, 1975; Karpicke & Roediger, 2008; Roediger & Karpicke; 2006a, 2006b). The spacing of testing of learning is known as the testing effect (e.g., Roediger & Karpicke, 2006b). *Retrieval* or the *generation effect* (Bjork, 1975) is the process of trying to remember what has been learned and recalling the learning. The two terms: retrieval practice and the testing effect are used to mean the same in this study: the benefits of retrieval practice.

Recently, Yang, Luo et al. (2021) conducted a meta-analysis of 222 papers about the testing effect to check for its robustness outside of the laboratory. This study is important to me as my study was conducted in primary classrooms. They suggest that testing gives a variety of benefits for learning for males and females even though these benefits may not be fully understood by students, educators and policy writers. A positive effect for learning was recorded in 82.9% of the studies surveyed and importantly worked at all levels of education including elementary school or primary school in the U.K.

The testing effect has been tested with various types of learning, for example: prose passages (Roediger & Karpicke, 2006b), lists of words (Karpicke et al., 2016), foreign language vocabulary words (Karpicke & Roediger, 2008) and can work in a backwards (consolidation of studied information) and forwards (learning of new information) way (Yang, Luo et al., 2021). Retrieval practice may be especially effective for students, who have low working memory (Agarwal et al., 2017; Yang, Luo et al., 2021). For a teacher, this finding suggests that using the testing effect is inclusive practice for students with low working memory.

In this study, retrieval practice is not used for subject content recall such as vocabulary or prose passages. It is used differently, to recall a learning strategy. The principle of retrieval practice is integral to my study in two ways: 1. It is used as a learning strategy in the intervention. 2. The testing effect *is* the 'subject content' in this study.

To teach the students about the testing effect, I used some of the research papers mentioned above and tried to recreate the ideas in Year 6 (10-to11- year-olds) friendly language. (see Appendix B). The idea of *study, test, test, test* being more effective than *study, test, study, test, study test,* or *study, study, study, test* (Roediger & Karpicke, 2006b) was promoted in the intervention and understood by the majority of students, who received the intervention (see Discussion and Appendix D).

As a teacher, I think one possible reason is because of the repetition used in these phrases; they are short, catchy and memorable. Students will often stop me in the corridor and say 'study, test, test, test, Mrs. Barsham' and my own daughters groan every time they hear the phrase! In my roles as a teacher and headteacher, I have used the phrase *study, test, test, test* like a mantra, in conversation, in assemblies, student mentoring sessions and in staff training.

Roediger, Agarwal et al. (2010, 2011) and Agarwal et al. (2010) suggest that, although it may seem counterintuitive to educators, education in school benefits from additional testing. This situation is because tests or quizzes, essays, exercises, and effortful retrieval of information are powerful mnemonic devices. Kornell and Bjork (2007) state that relatively few students viewed testing as a learning event with most students viewing testing as a type of monitoring of what is known. This knowledge ties in with the ideas and principles behind students and their often, incorrect *judgements of learning.* Students believe blocked practice (or repeated study) is better for learning than testing (e.g., Bjork et al., 2013) and students often use suboptimal learning strategies (e.g., Dirkx et al., 2019)

As a teacher, I think that there is a real need to educate students and teachers about the role of testing and to try and move away from the common perception that many students (and teachers) have, which is that a test comes at the end of a few weeks of learning, is graded by a teacher and that the results may go towards or be an overall grade or mark of achievement. As Yang, Luo et al. (2021, p.1) propose testing is not only an assessment *of* learning but also an assessment *for* learning".

It may be that this change in how tests are perceived is happening in classrooms now, as retrieval practice was included into the Initial Teacher Training Core Framework for the first time in the U.K in 2020. (see Discussion)

2.1.1 Retrieval Practice with 10-to 11-year-olds

Karpicke, Blunt and Sumeracki (2016) conducted an experiment in retrieval learning that was specifically directed at 10 year-olds. The results suggest that whether recall testing or recognition testing is used, retrieval-based learning works with this age-group.

In another experiment, Karpicke, Blunt, Smith and Karpicke (2014) adapted a recall, retrieval experiment, previously used with college students, for use with fourth graders (9-to10-year-olds) and found that the information retrieved was much less with this age-group. They concluded that the experiment had shown that retrieval practices had worked with the fourth graders but that there needed to be more scaffolding in the recall task. The students in my study are older than fourth graders but it is possible they may need some more scaffolding depending upon the individual.

Lipowski et al. (2013) suggest that first grade students (6-to7-year-olds) can benefit from the testing effect. Their study was conducted with first and third grade U.S. students. In addition, third grade students (8-to 9-year-olds) were aware of the benefits of the testing

effect too. However, this view is challenged by Aslan and Bäuml (2016) who suggest that the testing effect does not work with younger elementary aged children but does with older elementary aged children. Ritchie et al. (2013) suggest that retrieval practice enabled 8-to12-year-olds to recall more geographical facts than using mind mapping strategies.

Agarwal et al. (2014) in a study that included 11- year-olds, reports 92% of students feel retrieval practice benefits learning and importantly, for my study, that 72% of students felt that retrieval practice helped them with test anxiety.

The lack of research into the testing effect with primary aged children is highlighted by these few examples in the literature about the testing effect with this age-group and especially for the U.K.

It is important, for educational and research purposes, to know that retrieval has worked with 10-to 11-year-olds (e.g., Aslan & Bäuml, 2016; Karpicke et al., 2014, 2016; Lipowski et al., 2013) but there are not enough experiments to refine and define the contexts, the types of students and their educational needs.

My study may add to the literature around the use of the testing effect with U.K. primary aged children by using retrieval practice to reduce test anxiety and therefore potentially increasing well-being and *academic buoyancy* (e.g., Martin & Marsh, 2008) for this age-group. The idea of academic buoyancy is characterized by the 5C's: composure, commitment, coordination, confidence and control. The intervention in this study may help in composure, confidence and control. Academic buoyancy and student well-being are not measured in this study. Worry as a cognitive element of test anxiety and self-efficacy in test taking are measured but a reduction in worry or cognitive test anxiety is potentially linked to the idea of academic buoyancy or vice versa (e.g., Putwain et al., 2016) and well-being (Putwain et al., 2021; Steinmeyr et al., 2016).

2.1.1.1 Effort and Error. The idea of effortful retrieval being the key to remembering (Bjork, 1975) is an important one in terms of using the testing effect, effectively. Bjork (1994a) explains that self-testing is unappealing to many students because of the effort needed in comparison to re-reading but that this difficulty is beneficial for long term performance.

As a teacher, I have seen students cram information and knowledge into their brains, in preparation for a test. The image that springs to mind is of putting on a rucksack (backpack) of knowledge, that is taken off, as soon as the test is over. There are two problems

with studying for tests in this way. Firstly, the knowledge may not be retained beyond the test and secondly, because the students might achieve well in the test (cramming, or *blocked practice* works in the short term), they believe they know the information and do not attempt to re-study or test themselves on it again. This situation is called *retrieval fluency*. (Bjork et al., 2013)

In a testing situation, the idea of *retrieval fluency* manifests in learning content and then being tested directly afterwards and being able to recall the content effectively (Bjork et al., 2013). A student thinks they know the content, they achieved good or full marks. This will happen because the content is still in the short-term memory and is easily accessible. The problem occurs if you are tested on this information a week later and, in most cases, cannot recall the content at the same level as the test that took place immediately after the learning. Students may have a misconception of confidence in their ability. One way to manage this situation is to delay the initial retrieval of information. By delaying the first test, it makes the content harder to remember but more likely to be remembered or recalled a week later (e.g., Karpicke & Roediger, 2007a, 2007b)

One of the most effective forms of learning is achieved by using the testing effect because tests create retention of information more so than repeated study and even, in circumstances, where there is no feedback. (e.g., Roediger & Karpicke, 2006a). It has been shown that incorrect retrieval of information during tests is still a positive process (e.g., Kornell et al., 2015). Getting the answer right in tests is not as powerful learning as using and practising the recall process itself (Kornell et al., 2015). The very act of trying to remember is increasing the encoding of information going into the memory. Once, given the answer, the learning becomes elaborated and therefore easier to remember, because there are more links or cues to the information. Consider a scenario, ‘I remember Mr. Smith leaning over my desk with the pencil in his mouth and telling me that answer’ or ‘I remember getting that answer wrong last week’. Contextual differences in inputting information into memory are what make us human rather than machine. (Bjork, 2014) The more cues we can attach to knowledge to be learned, the better the recall will be. Indeed, making mistakes and getting it wrong are powerful mnemonics. (e.g., Bjork & Kroll, 2015) Corrective feedback, in the guise of Mr. Smith and other ways, is important in using the testing effect optimally and the role of feedback is discussed further in section 2.1.1.4.

The testing effect means that more learning occurs through being tested on information rather than studying information. It is the process of trying to remember, the

retrieval, that is powerful. To recall information rather than re-read or re-study information is harder. That is why the testing effect is a *desirable difficulty*. (Bjork, 1994)

Bjork and Kroll (2015) support the idea of effortful first retrieval and the evidence that initial challenges in retrieval benefit later retention and transfer of learning in vocabulary learning. The authors suggest that there are certain difficulties that are part of the process of learning a language. They propose that undesirable difficulties should not be created because desirable, in the case of language acquisition, is about responding to the difficulties successfully and by engaging in the processes that support learning, comprehension and remembering. They suggest that errors can promote learning and retrieval attempts that are incorrect can improve subsequent memory of the knowledge to be learned. Bjork and Kroll (2015) suggest that thinking of an incorrect answer gave the best later recall of a translation. Generating vocabulary incorrectly (Bjork & Kroll, 2015) and incorrect retrieval (Kornell et al., 2015) are examples of enriching cues and elaborating memory to enhance future recall. Getting it wrong is powerful learning and the types of difficulties that produce optimal learning will be different for different people.

As a teacher, I have always promoted making mistakes as a powerful learning event. I value mistakes to avoid a perfectionist culture in a school and to create resilience and academic buoyancy (e.g., Martin & Marsh, 2008). I have seen that perfectionism can be a problem for some students. As a headteacher (principal), I have led assemblies about famous people who made many famous mistakes and created reward systems in schools that celebrate making mistakes. However, some of the problems that I encountered came from the students being critical of themselves as a result of proximal and distal reasons such as: upbringing, teachers, previous school experiences, inherent characteristics and peer pressure. These factors can contribute to a fear of getting it wrong. In addition, the content driven nature of exams and performance monitoring of teachers can mean that teachers simply haven't got time for children to get it wrong and to learn from their mistakes.

Memory works in such a way that getting the learning wrong is one of the best things one can do. This idea of errors and mistakes being powerful learning is an aspect of many schools' development plans and I have personally promoted the idea in the schools I have worked in, citing the examples 'The Paris Festival of Errors' and San Francisco's 'Failcon' that are mentioned in 'Make It Stick' (Brown et al., 2014, p.93). However, schools still have league tables and performance data where individual schools are ranked according to performance against others. High stakes' tests that determine entrance to schools, universities

and sometimes school fee reductions have already been outlined as barriers to real learning, where making mistakes can be encouraged. The problem is that there is no time, and the stakes can be too high to allow for failure. Hence, cramming or blocking and regurgitating content for the test and forgetting it afterwards is often, necessarily, the modus operandi.

Learning about how encoding in memory works might help persuade students about the power of mistakes in learning, but school cultures need to change to value this kind of effortful and often slower but more effective, over the long-term, *desirable difficulty.* If the emphasis could remain on all round student outcomes rather than achievement, there could be a reduction in the number of students, who are high in test anxiety, and an increase in the number of students who learn how to learn for life, valuing tests for their ability to store information in the long-term memory, whilst equipping themselves with self-efficacy in independent learning strategies, changing the perception of testing.

As a contributor to a primer (Baker et al., 2020) for new P.G.C.E. (Post Graduate Certificate in Education) students at Cambridge last year, I was pleased to see that the new teaching standards recognize the importance of retrieval practice and also the benefits of making mistakes in learning.

From the Initial Teacher Training (ITT) Core Content Framework, 2020:

- Creating a positive environment, where making mistakes and learning from them and the need for effort and perseverance are part of the daily routine.
- Prior knowledge plays an important role in how students learn; committing some key facts to their long-term memory is likely to help students learn more complex ideas.
- An important factor in learning is memory, which can be thought of as comprising two elements: working memory and long-term memory.
- Regular purposeful practice of what has previously been taught can help consolidate material and help students remember what they have learned.
- Requiring pupils to retrieve information from memory and spacing practice so that pupils revisit ideas after a gap are also likely to strengthen recall.

As a teacher, I know it is important to allow students to make mistakes and learn from them. One of the best practices I have used in the classroom, is to make mistakes in front of the students myself; to role model getting it wrong. I have led CPD (Continuing Professional Development) in many schools and when training teachers and other stakeholders, I have

made deliberate mistakes in spelling and used techniques such as changing fonts to vary the input, to create cues for remembering. Putting one's hand up to point out a spelling mistake is effective learning. Many more cues for remembering are created through elaboration, rather than just reading the information, and are being used, as a hand is raised, and a mistake pointed out or a question asked.

However, the potential irony of the situation is that if you were to place the teaching standards next to teacher performance expectations in terms of student achievement grades, and the need to cover the content heavy, subject matter of many U.K. curricula, it may be difficult for teachers to make the two expectations work together.

2.1.1.1.1 An Aside About Effort. In an aside experiment, from this study, and not reported here due to the absence of a control group and limitations with small sample size, I had the opportunity to explore the theory of effortful retrieval and making mistakes with primary school children in the weekly spelling tests. After years of teaching English and spellings and children learning weekly (blocked) spellings for a weekly (blocked) test that still resulted in the correctly spelt spelling in the test, being spelt incorrectly in the context of writing in the classroom, I changed how spelling was taught in my school. I made this change in consultation with staff and parents.

Parents were invited to a talk about *desirable difficulties* (Bjork & Bjork, 2011) and told that the 20 out of 20 correct spellings' results may no longer apply. I explained that the learning would be "harder but stronger and last longer" (Brown et al., 2014, p. 9). Spelling lists for Year 1 to Year 6 were handed out, weekly as usual, but the testing interleaved the content. Weekly spelling tests combined different weeks' spellings. Corrective feedback was always given. There wasn't an opportunity to measure the results empirically but anecdotally, even repeated mistakes began to be spelt correctly after a few weeks, including a particular challenge for one of my students, 'astronaught'. This new spelling system took some monitoring, by the teacher, of the words tested in each week and it made learning spellings harder. The retrieval is harder but the elaboration greater because of the opportunities to make mistakes and receive feedback.

There would need to be further studies and a longitudinal evaluation to fully assess the effectiveness of this initiative and the impact for students who have dyslexia and other individual needs. Pan et al. (2021) recently published a paper asking if we still need to teach spelling in schools and they conclude that we should continue but that more research is

needed into using different strategies and known strategies with bigger sample sizes. My toe in the water trial at using desirable difficulties to teach spelling may evoke memories of the successful *Speller* scheme (Horn & Ashbaugh, 1920) that Pan et al. (2021) discuss for its use of distributed practice.

Moreover, as a headteacher, I was committed to the process of effortful learning and the promotion of making mistakes. Ultimately, the learning that was happening was that it was all right to make mistakes and learning to spell can take time but the learning lasts for a lifetime rather than a test. I do acknowledge that there isn't always time for making mistakes when having to deliver a content driven syllabus and I think that this issue is a barrier to real learning.

Another teaching spelling experiment I trialled was called 'Spelling Rocks' and involved students finding spellings and their corresponding rules (e.g., y = ies) in a type of spelling treasure hunt where spellings were hidden under rocks. This lesson was popular and was conducted outside. This system uses the idea of places to recall content (e.g., Imundo et al., 2020)

My ideas for a blog to highlight some of these problems and solutions with how spelling is currently taught is called 'Taughtology; the study of teaching spelling.' In this blog, I will include my ideas about changing how spelling is traditionally taught in schools through using desirable difficulties. The idea has been accepted by the learningscientists.org (http://www.learningscientists.org/).

2.1.1.2 Optimal Testing Effect Conditions. Roediger et al. (2010) suggest that the act of taking three tests stopped forgetting in its tracks. Cramming or blocked practice works if the need to remember is for short retention intervals. The image of putting on a rucksack (backpack) of knowledge, using it in an exam and then taking it off has been mentioned previously. Effortful retrieval is crucial for the testing effect (e.g., Butler & Roediger, 2007). The best formula is to recall through an initial short answer quiz and not through multiple-choice type tests. (Kang et al., 2007). However, when using pre-tests, multiple choice type answers may be more beneficial than pre-tests that feature cued recall. (Little & Bjork, 2016)

In addition, students need feedback that is corrective. (e.g., Roediger et al., 2010) The beneficial effect of testing can be altered if the initial test performance is low, and no corrective feedback is received. (Roediger et al., 2010). However, retrieval failure can enhance future encoding; not being able to remember or remembering incorrectly but then

getting corrective feedback, helps to encode the information into the memory in the long term. Delaying the feedback boosts final performance more so than immediate feedback, as long as it includes the correct information and the student understands (e.g., Metcalfe et al., 2009)

A key concept for the students, teachers, educators, school stakeholders to understand is that the timing of the first test of learning is crucial. Testing immediately after learning is testing the short-term memory and is not effortful. It is easy to remember the content at this stage. There is a real danger of *retrieval fluency* (Bjork et al., 2013). Learners run the risk of thinking they know the material because they were successful in a test straight after learning. The timing of the first test is key. Some forgetting of the studied material needs to have occurred so that the remembering is hard. (e.g., Bjork, 2011; Bjork, 2015; Bjork & Bjork, 2019) This forgetting will increase the encoding of the learning and increase the retention of the learned material (e.g., Roediger & Karpicke, 2011).

Whitten and Bjork (1977) attribute the powerful learning and retention that happens when an initial recall is delayed and made harder, to the memory searching through stored items to find links, matching cues in the semantic and acoustic memories. They conclude that the benefits of a hard, initial recall are twofold: firstly, searching for the information and finding it or being told the information afterwards, having searched, will strengthen the item's encoding power. This process will make it more durable and less prone to forgetting. Secondly, because the retrieval is taking more time and is not immediate, its journey is more complex, it is slower into memory and creates a more elaborate route into the memory and thus, will be more recallable. The premise is that quickly attained knowledge can be forgotten just as quickly. The putting on and taking off, of the rucksack (backpack) of knowledge. Karpicke and Roediger (2007a, 2007b) support this view saying that what is important for long term retention is the difficulty of the first retrieval attempt and that delaying the first test is critical.

The intervention in this study is not subject specific and does not test specific types or subject content in learning. The intervention used the general premise of retrieval practice to recall knowledge of what the testing effect is, itself. The intervention was delivered at a time when the students were taking numerous practice tests. The idea behind the intervention is that teaching 10-to11- year-olds about *their testing routes being well-oiled* (a phrase used in the intervention) in a quick, easy to deliver 15 to 20 minutes weekly session, over six weeks, would increase general feelings of self-efficacy about test taking.

The testing or retrieval, the recall of what could be remembered about the testing effect happened at weekly intervals. This schedule allowed time for forgetting about the testing effect and made retrieval more effortful the following week and therefore, hopefully, the knowledge more retainable. The principle of recall of the previous week's learning and then writing about the testing effect on two different occasions, over the six weeks is an example of repeated retrieval practice or the testing effect. According to Karpicke and Roediger (2007b, p.151) "repeated retrieval during learning is the key to long-term retention".

Ideally, the optimal conditions of testing outlined above (effortful, allowing some forgetting, study, test, test, test, or repeated retrieval) and those conditions outlined below (expanded vs. equal retrieval) could be used, alongside the type of intervention in this study, for maximum impact in terms of knowledge acquisition of subject content and simultaneously by teaching how the testing effect works, creating feelings of self-efficacy in test taking and a reduction in test anxiety. (Barsham & Ellefson, 2020).

2.1.1.3 Expanding Vs Equal Retrieval Practice. Roediger and Karpicke (2011) suggest that equal retrieval practice is better than expanding retrieval practice. In the expanding retrieval condition the first test occurs straightaway and that conflicts with what is known about the need for an effortful initial first retrieval. (Roediger & Karpicke, 2011) This situation is because a student can use the short-term memory to remember easily.

Roediger and Karpicke (2011) state that equal retrieval practice is better with the initial tests occurring a day or more after the initial learning and this condition is called equal retrieval practice. What is important is the difficulty of the first attempt at retrieval. They report that the equal spaced condition gives superior retention two days later and that delaying the first test improves long term retention. They recommend an interval of two days between recall events. The intervention in this study used repeated recall at weekly intervals as this suited most school timetables and may need adapting to feature two days recall depending upon timetables, teaching constraints and school contexts.

2.1.1.4 Feedback on Tests. Feedback is a priority in all aspects of education. Kornell, Hays and Bjork (2009) corroborated Izawa's (1970) findings which were that tests make the learning more powerful when the answer is presented after a test. Izawa (1970) searched for the number of tests that might produce optimal acquisition of learning and the answer was seven. Izawa's experiment notes the need for feedback in between testing as without

feedback one can keep testing up to a total of 19 times and neither learning nor forgetting occurs. Izawa's experiments showed that where testing answers are wrong and feedback is provided, the encoding of the learning is powerful to the extent that as much learning happens as when tests and study opportunities are interleaved.

Marsh, Agarwal and Roediger (2009) suggest that it is critical that feedback is provided under conditions in which it can be carefully processed. Roediger, Agarwal et al. (2010) refer to the direct benefit of testing material for powerful retention in memory and that providing feedback of the correct answers improves the testing effect. They comment that this feedback is very important when initial test performance is low. In addition, they suggest that multiple tests produce a larger testing effect and short answers, or essay tests produce a better testing effect than multiple choice or true /false options. Yang, Luo et al. (2021) meta-analyses of the testing effect found that corrective feedback gave more re-exposure to the learned material resulting in larger gains in learning. Butler et al. (2008) and Roediger and Marsh (2005) suggest that if students make errors on the initial tests and do not receive corrective feedback, then they can retain those errors on later tests. Delaying feedback often promotes better long-term retention. (e.g., Butler et al., 2008)

In this study, the feedback was not corrective on subject material, but it did happen on a weekly basis. Informal teacher and student conversations checked the understanding of desirable difficulties and especially the testing effect. The worksheets in the intervention ensured that the Year 6 teacher could quickly see what the students had understood. There was no formal requirement for the students to complete these weekly homework sheets, but some students did. These sheets alongside student thoughts were discussed briefly the following week, in class, and ideas added to or clarified. (see Appendix B).

2.1.2 Desirable Difficulties

In this section, I consider some of the literature around *desirable difficulties* (Bjork & Bjork, 2011). The purpose of reviewing this literature is to provide further understanding about how memory works. These ideas are important to understand when designing a new intervention that uses cognitive theory of retrieval practice, as its mechanism. This section helps to understand the issues around memory that led to the design of the intervention in this study that is aimed at increasing feelings of self-efficacy in taking tests, to reduce test anxiety in 10-to 11- year-olds.

Bjork and Bjork's (2011) desirable difficulties are changes in the learning context that create challenges for the learner but enhance long-term retention of the knowledge. The processes: varying the conditions of practice, for example the room used for studying (e.g., Bjork, 2011; Imundo et al., 2020; Smith, Glenberg & Bjork, 1978), spacing study or practice (tests), interleaving and not blocking and generating information retrieval, the testing effect are designed to improve encoding of information going into the memory, thus improving the retention of that information (e.g., Bjork, 2011). The more elaborate the encoding of information as it goes into the memory, the longer the knowledge will last (Bjork, 1975).

The desirable difficulties (Bjork & Bjork, 2011) have been examined by many experiments over the years and the results testify to the benefits of these processes to learning (e.g., Karpicke & Roediger, 2008). Bjork and Bjork (2011) credit spacing as having a great and significant impact on learning. Dunlosky et al. (2013) conducted a comprehensive survey on the effectiveness of ten different learning techniques. Typically observed student techniques such as re-reading, highlighting, keyword mnemonics, summarization and imagery use for text learning were found to have low utility whereas retrieval practice or testing and distributed (spaced) practice were found to have high utility. Interleaved practice had moderate utility in the research.

Information that is retrieved from memory becomes more recallable in the future, than it would have been, had it not been retrieved (Bjork, 1975). This idea is behind the intervention in this study when the students are told 'your testing routes are well-oiled'. This phrase or saying explains, in simple terms, the retrieval process and was used by teachers to try and instil a sense of self-belief, self-efficacy in taking tests in the 10-to 11-year-olds, who were involved in this experiment.

To my knowledge, this study is the first of its kind to use retrieval practice of the testing effect to try and create self-efficacy in test taking to try and reduce test anxiety with 10-to11- year-olds. (Barsham & Ellefson, 2020). At this age and stage in the U.K. educational system, the 10-to11- year-olds can sit tests in many subjects depending on the nature of the high stakes' tests and thus need test taking strategies, that can be transferred across different subject papers thus hopefully utilizing a perquisite of the testing effect that is transference of learning (e.g., Carpenter and deLosh, 2006; Carpenter, 2012; Kornell et al., 2009; McDaniel et al., 2013; Metcalfe, 2017; Pan & Rickard, 2018; Pyc & Rawson, 2010; Rickard & Pan, 2017) The intervention was applied universally and did not focus on a specific subject or type of learning, just on the testing process itself. The intervention focused

primarily on the testing effect. The other desirable difficulties were introduced in Week Two for extra information. The intervention is discussed in detail in Chapter 3 and evaluated in Chapter 10.

2.1.2.1 Judgements of Learning. A common problem that undermines the suggestions for improving long term retention of learning, is that many students are not able to judge their own learning (Bjork et al., 2013). Bjork et al. (2013) posit that learners' metacognition is unsuccessful in mirroring the benefits that successful retrieval gives to long-term retention. Karpicke, Butler and Roediger (2009) suggest that few students use self - testing to support their memories' performances. The perception in schools and colleges has historically been that testing is a means of assessing knowledge rather than learning (e.g., Bjork et al., 2013; Yang, Luo et al., 2021).

The lack of realistic understanding that students have about their own learning is discussed by Roediger and Karpicke (2006b) in an experiment about the meta-memory of prose passages. They found students' JOLs (judgements of learning) stated that repeated study was a better condition than repeated testing. Karpicke and Roediger (2008) report that students have no awareness of the mnemonic effects of learning. Bjork and Bjork (2019) discuss the benefits of interleaving over blocking or massed practice for perceptual motor skills (such as learning different baseball strokes), verbal-conceptual procedural skills (learning algebra) and the long-term retention of knowledge and transfer skills. The benefits of interleaving were evident in all types of learning. However, the students in this experiment did not believe in these benefits. They discuss how to get students to believe in interleaving. Despite being told that interleaving worked best and why it worked best, two thirds of the sample said they would still use a blocked method of teaching, if they were responsible for teaching that content.

Lipowski et al. (2013) suggest that third grade students (9-to 10-year-olds) were able to understand the benefits of the testing effect. My experiences as a researcher and as teacher in this study would confirm that the upper primary or elementary age-group understand the testing effect. (see Appendix D) It is possible that a younger age-group can understand aspects of cognitive theory such as the testing effect more readily than older students because they have used retrieval practice recently in classrooms and / or because they are not yet set in study habits such as blocked or massed practice. Moreover, it is possible that the educational climate is changing.

2.1.2.2 Blocking. The problem is that learners believe and feel blocking is better learning. Yan, Bjork & Bjork (2016) found that learners fail to appreciate three main things: the benefits of testing for learning, that making mistakes is good for learning, and spacing rather than massing practice or study leads to better long-term retention. Learners prefer non-optimal conditions for learning. Yan et al. (2016) investigated this problem further. They used a similar experiment to Kornell and Bjork (2008) that concluded that interleaving benefits are robust and effect size is large for benefiting inductive learning. The learners in Kornell and Bjork (2008) also reported that blocking not interleaving helped them learn better.

Yan et al. (2016), in a similar way to Bjork and Bjork (2019), attempted to mend the illusion that blocking is more effective for learning and this task was difficult. Although participants were informed about the theory of interleaving as opposed to blocking, this knowledge did not alter their preference. Yan et al. (2016), were interested in why this immoveable belief exists and concluded that it is largely a result of conditioning. Learners have been taught to believe that blocking is better and because it feels easier, blocking must be better.

In many cases, there may be students, who simply don't know how to learn in an optimal way. McNamara (2010) suggests this problem is a result of educational systems that are focused on delivering content but not how the content can be learned. This view is endorsed by Hattie and Yates (2014, p. 162) who call it a paradox that teachers expect students to remember accurately and note that there is little instruction about how to remember. However, it may be that the educational climate is changing. (see Discussion)

As a headteacher and senior leader in all through schools, I witnessed first-hand how colleagues struggled and continue to struggle to cover the content of the G.C.S.E.'s (General Certificate in Education) that were introduced in 2017. Indeed, many schools now start G.C.S.E.'s (General Certificate in Education) in Year 9 (13-to14- year-olds) so that they can cover the content over three years instead of two. As a headteacher, teacher (and mum) it concerns me that real learning techniques and a love of learning for life are being sacrificed because of the amount of content that has to be covered for a high stakes' test, that for many students becomes ten lots of blocked practice: a means to an end. The students need to get the grades to determine their sixth form studies (16-to 18- years-old). Most of the knowledge is not revisited afterwards unless further exams are taken in that subject. It is the G.C.S.E.

(General Certificate in Education) results that determine the 'A' level (Advanced General Certificate in Education) options or further education courses at 15-to 16- years-old.

As an educator of 24 years, I know that in many educational contexts in both primary (particularly upper primary as preparation for secondary education) and secondary phases of education, that the subjects studied are blocked. Many educational resources such as textbooks are compiled in a blocked way as are many examination syllabi. These exam syllabi and subsequent, necessary timetabling blocks ('blocks' are what the subject groupings in timetables are sometimes called in the U.K.) may contribute to blocked or massed learning techniques. However, increased use of retrieval practice and its recent inclusion into the Initial Teacher Training Core Framework (2020) may be changing the educational climate, especially at the upper primary level (see Discussion).

2.1.2.3 Retrieval Fluency. Another hindrance to judging learning effectively is discussed by Bjork et al. (2013) in the idea of retrieval fluency. This term means that the students believe they know the content because it is easily recallable. For example, reading a passage and then being tested on the material, immediately afterwards, will produce easy and non- effortful retrieval, that will make the student believe they know the material. Effortful retrieval is necessary and comes from delaying the initial recall of knowledge. (e.g., Roediger & Karpicke, 2011)

2.1.3 Possible Solutions for Improving Students' Metacognition

According to Kornell and Bjork (2009), there are two aspects to successful metacognition. Firstly, individuals need to know how memory works and secondly individuals need to have the ability to monitor one's own memory.

Koriat and Bjork (2006) found two procedures that helped to mend metacognitive thinking. Firstly, leading learners to use theory-based debiasing and secondly making learners aware of mnemonic cues around the ease of retrieval. These procedures provide people with valid knowledge of the cognitive system and encourage them to apply that knowledge in forming metacognitive learning strategies.

A test anxious student might become less test anxious if they are accurately and well informed about their learning journey. This knowledge could alleviate test anxiety. Knowing about how learning works and using this knowledge as a coping strategy to manage test anxiety, is the premise of the intervention in this study.

The ability to monitor learning accurately and to notice differences in the acquisition of learning and in testing and later testing is not an easy process. Monitoring of learning needs to be tracked and analysed and acted upon by students and teachers. Students need support in this process which is probably more important than many other educational processes (e.g., Bjork et al., 2013; Hattie & Yates, 2014). Bjork et al. (2013) refer to the issue of students rarely being tested on how they learn to learn. As an experienced educator, I would suggest that historically (presently?) students have not been taught how to learn or have been taught incorrectly. The intervention in this study does test students on how they learn for tests and teaches them that tests are learning events in themselves.

For a student, who is anxious about tests, to have a record of their learning during the learning process could feel very supportive and this self-knowledge could potentially be very empowering. A counter-theory is that for students, who are high in test anxiety, even the process of knowing they are studying for a test, the process of learning to learn for a test, can cause anxiety (e.g., Zeidner, 2014). However, in my opinion, if learning about how memory works was embedded in the curriculum, this may alleviate this problem for students who are high in test anxiety as it would become normal practice to monitor their knowledge and not learning associated with taking tests.

Pan and Bjork (in press) suggest that knowing how to learn is a critical survival tool. Bjork et al. (2013) regard metacognition as the ultimate survival tool. One problem is that unless the test has been taken after some forgetting has happened, the perception of knowledge may be subject to retrieval fluency (Bjork et al., 2013).

Agarwal et al. (2008) in their study about the testing effect with open and closed book testing, suggest that students failed to predict the effectiveness of testing relative to studying. However, Agarwal et al. (2014) suggest that 92% of middle and high school students felt that retrieval practice helped their learning and 72% said that retrieval practice made them less anxious about tests. Perhaps, this difference over time, reflects the way that classroom cultures are changing.

Learning about learning is a skill that can be taught (Pan & Bjork, in press). It is this idea that is at the heart of this study's intervention, because students were taught about how the testing effect works to give them the knowledge that their "testing routes had been well-oiled" prior to high stakes' tests. Knowing that the testing effect is a cognitive process that

promotes effective learning and retention of knowledge, has important implications for educational practice. (see Chapter 9, Discussion and Chapter 10, Reflections)

Tullis et al. (2012) recognise the importance of students having knowledge of the testing effect. Tullis et al. (2012) suggest that, with support, learners' knowledge of the testing effect can be improved. The students in my study were supported in learning about the testing effect. Moreover, for students, who are test anxious, this control over the learning could give feelings of self-mastery and self-efficacy.

As a headteacher and teacher, I think students should be taught the following key aspects about how to manage their learning: how memory works and specifically, the activities that help store knowledge in the memory, in other words retrieval practice and that retrieving information from memory is a memory modifier (Bjork, 1975); the role encoding plays in retrieval practice. (Bjork et al., 2013) and the learning biases of *hindsight*, *foresight,* and *stability* bias.

Hindsight bias (Fischoff, 1981) is when we think we knew the information all along and this misconception can then affect what we study or learn for an exam. *Foresight* bias (Koriat & Bjork, 2006) is when easy recall of information, straight after reading a text, does not mean the information is learned but the learner thinks they will be able to reproduce this information again easily. *Stability* bias, (Kornell & Bjork, 2009) is where students believe their memories will not change in the future and assume that what they know now, they will always know and be able to recall. Memory does not work in this way.

Knowledge is not stored like a tape that can be played at will but needs constant elaboration, association, cue recall and relating to previous information (Bjork, 1975). It is my opinion, as a teacher and educator, that the knowledge of these key aspects about how memory works and biases could transform what is learned in schools and the way it is learned.

DeWhinstanley and Bjork (2004) discovered that when learners recognise the benefits that testing provides, they may naturally engage in better encoding and processing strategies for studying items, during subsequent lists, to boost memory performance. Recognising the benefits that testing provides could increase confidence, self-efficacy in test-taking, exam performance and ultimately, perhaps feelings of well-being. I have measured self-efficacy in test taking and worry in this study but not performance or well-being.

To address these misconceptions about learning, it is important to teach students to monitor their learning accurately. An accurate appraisal of the learning process might give mastery of one's own learning and a key tenet of self -efficacy is mastery (Bandura, 1997). This strategy could be a key for a student, who has high levels of test anxiety. Learning about the testing effect, may lead to self-efficacy in test taking and a reduction in the worry or cognitive element of test anxiety.

2.1.3.3. Difficulties with Desirable Difficulties. The process of retrieval itself can affect other, to be retained, knowledge, in a process called retrieval induced forgetting (Anderson, Bjork & Bjork, 1994). In retrieval induced forgetting, learning that is processed with similar cues to the learning that needs to be retained, can be forgotten in the process of retrieval.

Forgetting information increases eventual storage strength and can benefit learning (Bjork & Bjork, 1992). Forgetting is the inability to retrieve information at the time that it is needed but it is not an irrevocable loss from long term memory. It is a loss of retrieval strength but not storage strength. This idea is explored more fully in Appendix C dedicated to the 'New Theory of Disuse' (Bjork & Bjork, 1992), and how to approach teaching this theory for possible use in an intervention or as part of a curriculum is suggested.

2.1.4 The Testing Effect and Test Anxiety

This section is brief because the literature is limited and especially relating to 10-to 11-year-olds. To my knowledge there is only one study that has considered the use of retrieval practice as a strategy to reduce test anxiety for this age-group.

Agarwal et al. (2014) surveyed 1408 U.S. middle and high school students (11-to15-year-olds) and the results suggest that 72% of middle and high school students felt that retrieval practice helped them in exams. They suggest classroom teachers should adopt retrieval practice to manage test anxiety.

This research by Agarwal et al. (2014) has been featured on the Effortful Educator blog (https://theeffortfuleducator.com/2019/05/14/retrieval-practices-impact-on-test-anxiety-and-stress/). Although, retrieval practice is now included in the Initial Teacher Training Core Framework (2020) and on websites such as Learning Scientists.Org (https://www.learningscientists.org/) and Agarwal's retrievalpractice.org (https://www.retrievalpractice.org), there is very limited literature about the links between

using retrieval practice and its role in reducing test anxiety other than Agarwal et al. (2014) and the related blog from the Effortful Educator (2019).

Another study that has explored the relationship between test anxiety and the testing effect is Yang, Bukuan et al. (2020). However, the study is with university aged students and engages low stakes rather than high stakes' tests. The results suggest that the testing effect is an effective technique to enhance learning for individuals with different levels of working memory capacity and therefore supports the findings in Agarwal et al. (2017). In addition, Yang, Bukuan et al. (2020) suggest that frequent tests have minimal influence on test anxiety and is therefore a different result to Agarwal et al. (2014) where the authors found that 72% of middle and high school students felt that retrieval practice helped them with tests. The students were younger in Agarwal et al. (2014).

A quick Google search in March 2021 for 'the testing effect in elementary or primary classrooms' produced some research papers about retrieval practice. However, it is interesting that after a handful of papers the entries move quickly into the stresses related to testing. Yet, to my knowledge, there is only one research paper that has looked at Year 6 (and then only as part of a much larger sample) and using retrieval practice to reduce test anxiety (Agarwal et al., 2014) until now (Barsham & Ellefson, 2020).

My study takes the idea of using retrieval practice for reducing test anxiety into the realms of the pastoral as well as an academic curriculum in school. Teaching the testing effect using the testing effect at a time when students were completing practice tests anyway uses retrieval practice to give students a strategy to increase feelings of self-efficacy in test taking. My study is unique in this respect and may provide the intervention for Year 6 that Putwain (2020, p.6) has described as a "pressing need".

2.1.4.2 Does Retrieval Practice Work for Students who are High in Test Anxiety?

For an explanation of what high test anxiety is (see 2.2.3.2). To my knowledge, there is no research about using retrieval practice with students who are high in test anxiety in Year 6 (10-to11- year-olds). Hembree's heavily cited meta-analysis (Hembree, 1988) suggests that test anxiety can start at age seven. Putwain and Daly (2014) suggest between 10 and 20% of students in a classroom are high in test anxiety.

As a headteacher in a Prep (primary) school, I have seen that students of this age do experience test anxiety. As a teacher of Year 5 and Year 6 (9-to 11- year-olds) I have taught

students who are high in test anxiety in this age-group. This section will draw upon research with older students as that is all that is available.

Agarwal et al. (2014) suggest that retrieval practice made 72% of their middle to high school students feel less anxious about taking unit tests that counted for a grade. In addition, 81% of students said they felt the same or less test anxiety in classes where they had received clicker quizzes. However, 19% of students said they felt more anxious about tests because they had been using quizzing procedures. Students who have high test anxiety are not commented on in this isolated example of an age relevant test anxiety and the testing effect study, but Agarwal et al. (2014) suggest that future studies should focus on individual characteristics and the inducement of test anxiety from quizzing. It is possible that some of the 19% of participants in Agarwal et al. (2014) can be accounted for by students who experience high test anxiety.

Hinze and Rappe (2014) question the benefits of retrieval-based practice on highly test anxious, college aged students by stating that anxiety affects the cognitive process. In their *retrieval disruption* hypothesis, the authors examine the effects of low stakes' tests and high stakes' tests on students with test anxiety. The conclusion is inconclusive. The power of retrieval is clear and testified by the authors. The authors comment that there is a need for further empirical investigation as their experiment that offered financial rewards, in an attempt to create a high stakes' situation is not realistic of a truly evaluative situation.

Tse and Pu (2012) asked whether repeated testing is equally effective with all students. Their experiment tested college age students with lower working memory capacity and trait test anxiety. Once again, the benefits of testing were clear. Participants in the experiment performed better in the repeated testing condition than in a repeated study condition. However, the participants with a lower working memory capacity and higher test anxiety, made more intralist errors during the acquisition phase of learning and the testing effect was negatively correlated for this group.

Both Hinze and Rappe (2014) and Tse and Pu (2012) present different findings to Agarwal et al. (2014) that suggest 72 % of middle to high school students experience reduced test anxiety because they have used retrieval practice in the classroom. However, both experiments are with older, university aged students and conducted in laboratory conditions in contrast to Agarwal et al. (2014). Clark et al. (2018) support Aggarwal et al. (2014) and

challenge Tse and Pu (2012) by suggesting that testing is more beneficial with students who are high in test anxiety but again the participants are college aged U.S. students.

A student, who has high levels of test anxiety, may not perform as well in evaluative situations as a student, who has lower levels of test anxiety because students, who have high test anxiety, are less likely to perform the elaborate rehearsal that helps the encoding of information. (Carpenter, 2009). There will be limited encoding and cueing because anxiety is dominating, and all the transactional *mediating* and *moderating* variables are competing for attention. (Zeidner 1998). It is likely that an individual, who is test anxious, is busy with worrying and is less likely to come up with retrieval cues. This information emphasizes the potential need for an effective intervention for a student, who has test anxiety, at an early stage of the evaluative process, before maladaptive coping (Zeidner & Matthews, 2005) manifests.

However, there are the students who are high in test anxiety who use increased effort (Putwain & Symes, 2018) to compensate for performance deficits and thus the relationship between test anxiety and performance deficits is relatively low at r = -.21 (Seipp, 1991)

Students, who are high in test anxiety, can perceive tests as opportunities for catastrophic failure. McDaniel et al. (2013) suggest that it is possible that the benefits of testing may counteract the anxiety that is created by the process of testing, for a student who does experience trait and therefore state test anxiety because some students become less anxious after regularly being tested in class. Cassady (2004) says that students, who are high in test anxiety, showed low self-confidence in the test preparation phase.

However, Agarwal et al., (2014) in the only study, to my knowledge that does include 10-to 11- year-olds suggest that 92% of students felt that retrieval practice helped them learn and 72% felt that retrieval practice reduced test anxiety around unit tests and exams. This study examined the use of low stake quizzes using clickers in the classroom to enhance subject knowledge recall. As already stated, there is no specific information about how the students who are high in test anxiety felt about the use of retrieval practice in reducing test anxiety.

My study did not use subject specific retrieval but retrieval of retrieval practice. The testing effect on the testing effect. Each week, over a period of six weeks, the students were invited to discuss, to recall what they could remember about desirable difficulties and especially the testing effect. Teaching the testing effect, in this way, through recall has never

been done before with 10-to11-year-olds with the purpose of increasing test-taking self-efficacy to reduce test anxiety. Eighteen percent of the students in this study were high in test anxiety, $N = 31$.

Students need self-efficacy; the ability to believe in their capabilities. They need support in their academic self-concept (e.g., Lohbeck et al., 2016). Students can benefit from understanding how learning occurs. Knowing that testing is increasing long-term retention could give feelings of confidence when faced with an evaluative situation. Knowing that testing has been practised and that the "testing routes are well oiled"; the memory has been trained in testing, could give confidence to students, who are test anxious, in evaluative situations.

In my study, I hypothesize that knowing about the testing effect may be an appropriate way for creating feelings of security and self-efficacy in test taking through the ability to be able to recall this reassurance for students, who suffer from test anxiety and for those students who don't suffer from test anxiety, and for those students who experience test anxiety at varying degrees. Sharing the information from Bjork's (1975) *memory modifier* (retrieval makes it easier to retrieve) and the significant results from Roediger and Karpicke (2006a, 2006b) about the benefits of study, test, test, test plus the idea of non-successful retrieval (Kornell et al., 2015), where unsuccessful retrieval still aids future retrieval and learning, are possibly, cognitively powerful reassurances for students, who are anxious about testing and for other students who have varying degrees of test anxiety. This knowledge may build *cognitive morale.*

2.1.5 Summary and Next Section

This chapter has discussed the testing effect and the benefits to learning from using retrieval practice. This chapter has highlighted the lack of research into the testing effect with 10-to 11-year-olds through citing the few relevant studies (Aslan and Bäuml, 2016; Karpicke et al., 2014, 2016; Lipowski et al., 2013; Ritchie et al., 2013). This chapter has highlighted the lack of research for the testing effect to reduce test anxiety with 10-to 11-year-olds (Agarwal et al., 2014) and the absence of research for using the testing effect with 10-to 11-year-olds, who are high in test anxiety.

This chapter has discussed how the educational climate may be changing as demonstrated by recent changes to the Initial Teacher Training core framework, (2020) that

now incorporates retrieval practice and the idea of effortful and use of errors in learning into new teacher training.

The problem of students' metacognition in terms of their judgements of learning has been explored and as a researcher, I have considered some solutions to the problem and as a teacher, offered some further solutions.

I highlight the issue of mistakes in learning needing to be celebrated and time given for effortful learning rather than simply getting through the content of a syllabus. Perhaps the way students are assessed needs to be changed to allow for less content and more, real learning? Misconceptions in Science learning are common (e.g., Tsapali, 2020) and slowing down and making more mistakes that can then be corrected along the learning journey may reduce some of these misconceptions.

From this chapter, some ideas that are central to the intervention in this study emerge: the power of the testing effect, taught to the students through "study, test, test, test" (Roediger & Karpicke, 2006b) and the idea of memory being a modifier (Bjork, 1975) "your testing routes are well-oiled".

The next section examines test anxiety theory and constructs. It is necessary because to design an intervention using the testing effect to reduce the worry or cognitive element of test anxiety, the dynamics of the text anxiety construct must be understood.

A key feature of transactional test anxiety is an individual's perception of self-control. The intervention was designed using the testing effect to increase feelings of self-control, specifically self-efficacy in test taking. The concepts of self-control and self-efficacy are inextricably woven into the S-REF (Self-referent executive function) model of test anxiety (Zeidner & Matthews, 2005) in the self-knowledge beliefs strand of the model.

Section three of this chapter examines self-efficacy and in particular, the relationship between uncertain control (Putwain & Pescod, 2018), perceived control (Putwain & Aveyard, 2018) and test anxiety and inevitably ideas overlap with ideas discussed in the next section, section two that is about test anxiety and also ideas in section one, the testing effect. However, section three of this chapter has been designed to be read on its own so that the complexity of the interleaving concepts can be fully explored. The next section, section two of this chapter, is about test anxiety.

2.2 Test Anxiety

This section of the literature review discusses many aspects of test anxiety. It was necessary to understand how test anxiety works, from looking at the history of the construct, its various models, some related theories, previous interventions, and instruments of measure, to be able to create an intervention using the cognitive theory of the testing effect, to address the problem of test anxiety in 10-to 11- year-olds. The new and adapted measures used in this study are also informed from learning about test anxiety, the testing effect and self-efficacy theories.

In the DSM-IV, Diagnostic and Statistical Manual of Mental Disorders (American Psychiatric Association, 2000) test anxiety is not coded as a separate diagnosis but if it interferes with daily living because of its distressing levels, it could be classified as a social phobia. The ICD-10, International Classification of Diseases, (World Health Organisation, 1992) classify examination phobia as a specific phobia. There is a lack of clear diagnostic criteria for test anxiety (Bögels et al., 2020; LeBeau et al., 2020).

Zeidner (1998) investigates the origins and development of test anxiety and organizes the antecedents of test anxiety into proximal and distal categories. Distal factors are those that are more distant from the individual. Distal factors are a result of genetic make-up and / or upbringing. Distal factors are those that are not products of the immediate environment. One major distal factor is parental attitudes towards raising children. I differentiate parental input around the topic of exams during exam periods to a proximal factor.

As a headteacher, I have talked regularly to parents and teachers about the need to adopt an ‘it’s ok to make mistakes’ culture at home and school; to be mindful about the pressures placed on children to succeed. I have talked about the types of parental, teacher and peer transactions that can cultivate anxiety, especially around tests. If parents and teachers become anxious for the children as important tests approach, then, akin to the process of osmosis, the children can become more anxious too. I have appealed for mindful approaches, a sense of keeping perspective as the key testing points of an academic calendar are approached.

Proximal factors are those of the immediate environment. Zeidner (2014) highlights the school environment as a major proximal factor. These can be simply the physical context of an examination setting but also the result of many peer and social pressures; not least the

teacher or parental pressure that can be conveyed in *fear appeals* (e,g., Putwain & Best, 2011; Putwain et al., 2014, 2016).

Fear appeals are statements issued by teachers (peers / parents) to students prior to high stakes' tests and are designed to alert students to the importance of the upcoming test and the need to study hard, to prepare. Reminders of what is at stake such as university or school entry, scholarships or academic records may be used. Fear appeals may be meant to be helpful in motivating study prior to examinations but they may also affect self-belief. Putwain et al. (2014) suggest that for students, who had high test anxiety, fear appeals made them more anxious. However, for students, who have low test anxiety, fear appeals can work to make them study harder. In any given classroom, there are approximately 10% - 20% of the class that experience high test anxiety (Putwain & Daly, 2014).

Putwain (2018) discusses the antecedents of test anxiety as: beliefs in ability to achieve, study habits, skills in studying and test taking, how important a particular exam is and the evaluation of the exam and context as it happens and a person's approach to coping. Putwain (2018) suggests that, when looked at separately, the relevance of some test anxiety antecedents may become overstated and that studies have not yet examined the role that regulating or controlling the emotions plays in the transactional nature of test anxiety. One aim of this study is to try and contribute to this gap in the literature through trying to measure self-efficacy in test taking in 10-to11- year-olds, using pre-test and post-test timepoints around high stakes' tests.

Test anxiety affects performance in tests. (e.g., Hembree,1988; Putwain, 2008) Test anxiety affects the cognitive process and working memory (e.g., Eysenck & Calvo, 1992). There are various estimates for the number of test anxious children in a classroom (e.g., Putwain & Daly, 2014) and it is difficult to establish an exact measure (Cizek & Burg, 2006). Test anxiety is a global phenomenon (e.g., Zeidner, 2014). Test anxiety affects girls more than boys (e.g., Hembree, 1988). Test anxiety increases in response to high stakes' tests compared to classroom tests (Segool et al., 2013).

Gender has not been a focus in the research questions and analyses in this study. There are meta-analytic studies (e.g., Hembree, 1988) that have established that test anxiety affects girls more than boys. The main focus of this study is in trying to reduce worry about exams for 10-to 11-year-olds.

The construct of test anxiety is a complex and multi-dimensional one (e.g., Chapell et al., 2005; Wine 1982; Zeidner, 1998). Test anxiety can involve combinations: metacognition, strategies for coping, motivation, strategies for study, perceived importance of a test and attentional bias. (e.g., Putwain & Pescod, 2018)

Putwain and Symes (2018) describe test anxiety as a perceived threat to goals including educational or occupational hopes for the future and feelings about self-worth. Putwain, Chamberlaine, et al. (2014) suggest that evaluations of self-worth are the result of negative self- judgements that are a result of perceived failure. They outline the construct of test anxiety, as consisting of four elements: cognitive or worry, behavioural, affective, and social elements. The focus of this study is the cognitive element of test anxiety, worry.

Goonan (2003) suggests that test anxiety is more than normal worry about a test. It is a specific anxiety disorder that involves extreme amounts of concern, worry and fear about being evaluated negatively and manifests in the evaluative situation as difficulty in thinking clearly and being able to retrieve information. One should note that the conclusions in this study are not drawn from empirical data (Putwain & Pescod, 2018).

Zeidner (2014) discusses the ideas that intellectual, motivational, and social capabilities have to work harder, in individuals, who are test anxious. There is cognitive worry about failure, remembered from incidents in the past and fear of future failure and the negative consequences that will accompany such a failure. As a result of these thoughts, there is a reduction in the ability to be able to process information and respond accordingly in tests (e.g., Cassady, 2004; Zeidner, 2007). Furthermore, emotions increase and can result in heightened physiological responses such as increased heart rate, sweating and ‘butterflies’ (e.g., Putwain, Chamberlaine et al., 2014)

Test anxiety is associated with negative thinking and behaviour at all phases of the learning and testing cycle. It is not just a phenomenon that happens in the testing room but during the accumulation of knowledge and the also post-tests, when reflection occurs. (Zeidner, 1998). In a rare experiment, conducted in an actual exam, Galassi et al. (1981) suggest that for students, who are high in test anxiety, the performance aspect, sitting the exam, is the worst, negative and most intense part of the evaluative process.

Test anxiety is a multi-dimensional, multi determined phenomenon and a variety of different test anxious individuals can be identified. (e.g., Zeidner, 1998)

Zeidner and Matthews (2005) regard test anxiety as being a characteristic of the person and part of a dynamic process. Essentially, it is a complex construct, and each person is individually complex. Zeidner (1998) refers to the reciprocal interaction of several clear elements: the evaluative context, traits or characteristics of anxiety, the perception of the threat and how this information is assessed by an individual, state anxiety caused by the room, the instructor, previous failures combined with trait or characteristic anxiety and the ability to adapt and cope in the evaluative situation. Spielberger et al. (1976) define test anxiety as an individual's tendency to worry and to have irrelevant, intrusive thoughts and an inability to think straight and physiological arousal.

From the ideas outlined above, it is hoped that the reader gains a sense of the transactional nature of test anxiety. It is dependent upon the person, behaviours, and the environment. The test anxiety construct has a similar structure to the triad of reciprocal causation (Bandura, 1986b) of behaviour, personal factors, and environment: B = behaviour, P = internal personal factors such as cognitive, affective, and biological events and E = the environment.

2.2.1 The Test Anxiety Construct

This section examines the different models of test anxiety to afford the reader with a sense of how the model has changed. The history of test anxiety discusses drive models of test anxiety, deficit models, cognitive motivational models, and transactional models (Zeidner, 1998). Theories such as: expectancy, attentional, motivation and arousal (Zeidner, 1998) cross over into test anxiety theory but it is not within the scope of this thesis to discuss these in detail and thus they are considered only where relevant.

Liebert and Morris (1967) introduce the idea that test anxiety is a bi-dimensional construct defined by *worry* (cognition) and *emotionality* (behaviour). The worry element of test anxiety is the focus of this thesis.

Alpert and Haber (1960) refer to the *facilitative* and *debilitative* properties of test anxiety. Test anxiety as a facilitative concept is considered only briefly in this thesis because the idea of test anxiety being facilitative is over simplified. Putwain et al. (2020, p.3) explain that there is a difference between the facilitating effects of challenge and that this feeling of 'in the flow' (e.g., Csikszentmihalyi, 2004) should not be confused with the debilitating effects of anxiety arising from threat. They reject the argument of test anxiety being facilitative due to the absence of empirical support for the Yerkes-Dodson law (Yerkes-Dodson, 1908).

Sarason (1972) and Wine (1971) propose a *cognitive attentional* model. Carver and Scheier (1981) refer to people who set goals and standards. Tobias (1985) proposes a *learning deficit* model but notes that the problem with this idea is that students, with good study habits, can get anxious too. Covington (1992) suggests that individuals do have efficient study skills but that with test anxiety, they get blockages that they can't remove. Smith, Arnkoff and Wright (1990) suggest a *dual deficit* model with *interference* and *skills deficits*.

Naveh-Benjamin (1991) proposes the *deficient in study taking skills'* model that also features problems in being able to encode knowledge in study and then being able to retrieve this knowledge subsequently. Spielberger and Vagg (1995) define test anxiety as *A state* anxiety, which happens more readily for people who have *A trait* anxiety. Some people are naturally more anxious than others, as a result of many proximal and distal factors (Zeidner, 1998), and are more likely to become anxious in evaluative situations developing *A state* anxiety if they have *A trait* anxiety.

Zeidner and Matthews (2005) extend this model by describing integrating, multiple influences of anxiety on cognitive processes in the S-REF (Self-referent Executive Function) model of test anxiety. It is a strand of the S-REF (Self-referent Executive Function) model that is targeted by the intervention in this study. Lowe et al. (2008) introduce the biopsychosocial model, a combination of cognitive, learning, demographic, social and environmental influences.

Zeidner and Matthews (2005) discuss how a repeated history of failures makes individuals accept failure readily and that they can adopt a type of learned helplessness. They discuss different test anxious types: self-handicappers who may not study because they can then blame their failure on this process rather than anything else such as inability and, on the other end of the scale, are the perfectionist over-strivers, who ask too much of themselves.

In addition, Zeidner and Matthews (2005) broaden the understanding of the test anxiety construct through considering the theories around evaluation anxiety. It helps to understand the complex nature of test anxiety by considering multiple factors such as: negative self-belief, skills deficits, avoidance motivation, self-regulative strategies, maladaptive stress processes and drive/arousal theories. The latter of these is dismissed due to its failure to differentiate between emotionality and worry. (Zeidner & Matthews, 2005).

These multiple contexts and factors emphasize the transactional nature of test anxiety and can be an indication as to the problems of measuring and intervening in test anxiety.

Furthermore, Zeidner (2014) proposes *interactional personal traits* plus *situational anxiety*, which is similar to and a development of, through the word 'interactional', the *A trait* and *A state* model proposed by Spielberger and Vagg (1995). The social learning model of test anxiety (Smith et al., 1990) is discussed at 2.3.1.1 as the ideas in this model are linked to self-efficacy theory.

Test anxiety is a complex construct and research has been limited by the self-report aspect of this construct (Putwain, 2018) and because test anxiety is difficult to research as it happens in assessment situations (e.g., Schutz et al., 2014). Most test anxiety research is necessarily conducted pre and post evaluative situations. An exception is the study by Galassi et al. (1981) that was conducted in an actual exam.

The S-REF (Self-referent executive function) model (Zeidner & Matthews, 2005) is explored in much more detail as it is this model that is targeted by the intervention in my study. (see 2.2.1.2) A further key idea to consider is that although the research suggests that cognitive test anxiety affects performance in high stakes' tests due to its debilitating effects on working memory (e.g., Owens et al., 2008; Putwain, Shah & Lewis, 2014), counter theories suggest that students with high test anxiety compensate through extra effort to reduce the performance deficit (e.g., Eysenck & Calvo, 1992; Eysenck et al., 2007; Putwain & Symes, 2018) and retrieval practice works with students with low working memory (Agarwal et al., 2017) and moreover, that retrieval practice reduces test anxiety (Agarwal et al. 2014).

2.2.1.1 Worry. This study focusses on the cognitive or worry element of test anxiety as identified in the bi-dimensional construct of Liebert and Morris (1967). The focus on this element of the test anxiety construct and not the emotional (behaviour) or affective (physiological) elements is because the worry component of test anxiety is a wise target (Putwain & Pescod, 2018) for intervention because of its links to achievement (e.g., Hembree, 1988) and well-being (Putwain et al., 2021; Steinmeyr et al., 2016). In addition, the affective physiological elements of the test anxiety construct can be equivocal (Putwain, 2018) and it is difficult to identify specific behaviours that are purely indicative of test anxiety (Putwain et al., 2020).

The intervention in this study was designed to try and reduce the worry element (Liebert & Morris,1967) of test anxiety using the cognitive theory of the testing effect,

through trying to increase feelings of self-efficacy in test taking by intervening in the self-knowledge beliefs strand of the S-REF (Self-referent Executive Function) test anxiety model (Zeidner & Matthews, 2005).

Liebert and Morris (1967) shifted the emphasis of test anxiety into the cognitive realm through their work, (Zeidner, 2014). It is the cognitive element that affects working memory and performance in assessments more so than *emotionality*, also referred to as *behaviour*. (Eysenck & Calvo 1992; Hembree, 1988; Liebert & Morris, 1967; Putwain, 2018)

Zeidner and Matthews (2005) summarise test anxiety as a three-dimensional construct: *behavioural*, *cognitive,* and *affective-physiological*. The test anxiety construct, that the Children's Test Anxiety Scale (Wren & Benson 2004) seems well aligned to measure. In the Children's Test Anxiety Scale, there are three subscales: *thoughts*, that measures the cognitive or *worry* element of test anxiety, *off task behaviours,* that measures behaviour and *autonomic reactions,* that measures the affective physiological element of test anxiety. The Children's Test Anxiety Scale (Wren & Benson, 2004) is the measurement that I chose to use for the pre-test and post-test questionnaires in the pilot and main study.

Putwain & Aveyard (2018) address *perceived control* and test anxiety. They examine how three self-beliefs (academic buoyancy, perceived control, and test competence) can work to decrease worry, or the cognitive element of test anxiety, and how this decrease in worry increases perceived self-control and examination performance. They suggest that perceived control, but not academic buoyancy or test competence moderated the relationship between worry and test anxiety and that some students with high perceived control can still become anxious about tests. For example, if they hold beliefs about the usefulness of worry. (Putwain & Aveyard, 2018) Uncertain control and perceived control in test anxiety are found in the self-knowledge beliefs strand of the S-REF model (Zeidner & Matthews, 2005) of test anxiety (Putwain, 2018) and it is this layer of the S-REF model that is the target for the intervention that aims to increase self-belief, self-efficacy in test taking.

In my study, I focus on the idea that self-belief, the belief in knowing about how the testing effect works can increase feelings of self-efficacy about taking tests to reduce worry or the cognitive element of test anxiety thereby increasing perceived control and reducing uncertain control. This idea of increasing self-control to reduce worry or cognitive test anxiety is key to this study and explored further at 2.3.2 and 2.3.2.2.

2.2.1.1.1 Other Test Variables. Zeidner (1998) discusses the variables that need to be considered for a student who is test anxious. They are listed here with some specific examples included from my teaching experience, to provide further context about contributing factors to test anxiety: 1. How hard a test is may affect different students in different ways; 2. The order of the questions. As a G.C.S.E (General Certificate of Secondary Education) examiner for many years, the paper usually started with easier questions that were designed to 'warm up' the students in preparation for harder questions later on in the paper; 3. The test format. As a teacher, I taught exam rubric specifically so that there were no surprises for students about what the exam looks like and when delivering mock exams ensured that the experience was identical to the real exam even down to the type of booklet used for writing; 4. The choice in the items to answer. It is not always possible to provide an option, but many exams do feature an element of choice at some point. For example, in English exams, the comprehension is usually compulsory and will be followed by a choice of essay questions. As a headteacher, I used InCAS (Interactive Computerized Assessment Systems) to assess primary aged students as this computerized system differentiated questions according to questions answered.

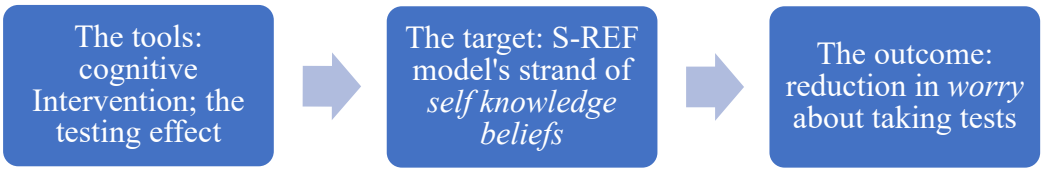
In addition, the exam room, the environment itself can cause test anxiety. Furthermore, the instructor may not be known to students in an exam and for some students this reality can make them feel anxious. In designing a questionnaire, I tried to measure some of these variables, I included questions about the instructor and the exam room and the degree of difficulty of the test (see Appendix A; Appendix H, Figure H.1; Figure H.2).

2.2.1.2 The S-REF Model. Section 2.2.1 outlined some of the test anxiety constructs, but it is important to explore the S-REF (Self-referent Executive Function) model by Zeidner and Matthews (2005) in more detail. The reason is because it is the self-knowledge beliefs' strand that is targeted by the intervention in my study, that aims to increase self-efficacy in test taking and reduce test anxiety in 10-to 11-year-olds.

I have been led to this target point from reviewing the literature. A key element to focus on came from the worry and emotionality model of test anxiety (Liebert & Morris, 1967) and the idea that it is the worry or cognitive process that affects performance but more importantly the cognitive element or worry that can alter perceived control (Putwain & Aveyard, 2018) and reduce uncertain control (Putwain & Pescod, 2018) and potentially improve student feelings of self-efficacy that may manifest as improved beliefs about test taking capabilities. (see Figure 2.1)

Figure 2.1

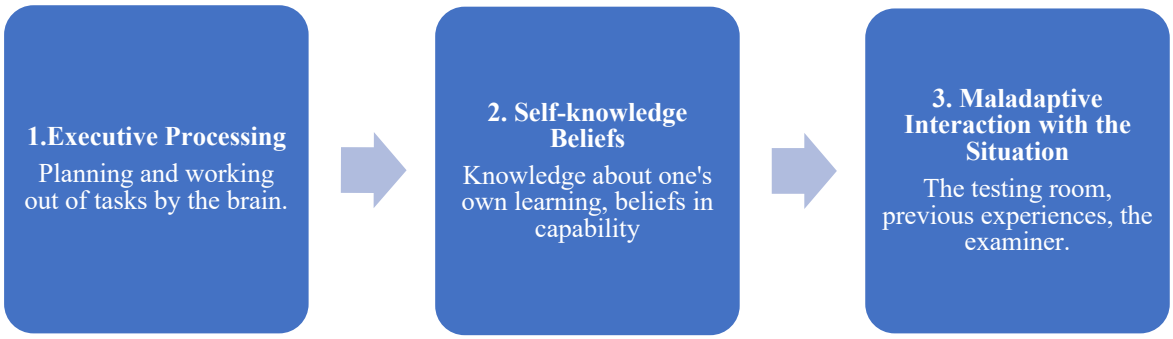
The Mechanism



The S-REF (Self-referent Executive Function) model of test anxiety (Zeidner & Matthews, 2005) summarises three areas that illustrate the test anxiety process (see Figure 2.2)

Figure 2.2

The S-REF Model - Self-referent Executive Function (Zeidner & Matthews, 2005)



The S-REF (Self-referent Executive Function) model of test anxiety (Zeidner & Matthews, 2005) spreads across several cognitive processes. An individual will receive notification of an exam and assess the degree of threat through accessing self-knowledge beliefs; for example, about previous successes or failures in evaluative conditions, beliefs in test taking capabilities. After that, decisions about how the individual will respond are made. It is at this point, that maladaptive coping strategies, such as mental or behavioural avoidance of the evaluative situation or turning to maladaptive metacognitive beliefs happens. For example, 'If I worry, it means that I won't fail' types of self-talk. (e.g., Zeidner & Matthews, 2005). This experiment is designed so that the intervention intervenes at the self-knowledge beliefs point of the S-REF (Self-referent Executive Function) model (Zeidner & Matthews, 2005) to try and improve feelings of self-efficacy about taking tests, to reduce the worry element of test anxiety.

2.2.2 The Measures

von der Embse et al. (2020) suggest that many existing test anxiety measures are limited because they fail to reflect developments in the test anxiety phenomena over time. In this study, I chose a measure that had been used with children of the same age-group, but I explored it and adapted it to reflect possible changes that may have occurred in the test anxiety construct over time. The Children's Test Anxiety Scale (Wren & Benson, 2004) was the established measure that I chose. A recent comment by von der Embse et al. (2020) suggests that this measure does not consider students' perceptions of test use. In my study, I created a new measure to assess self-efficacy in test taking that may capture some students' perceptions of test use.

There are a few measures for test anxiety that have been developed over the years for children: The TASC, Test Anxiety Scale for Children, (Sarason et al., 1958); TAICA, Test Anxiety inventory for children and adolescents (Lowe et al., 2008); STAIC, State Trait Anxiety Inventory for Children, Spielberger et al. (1973); CTAS, Children's test Anxiety Scale (Wren & Benson, 2004); MTAS, multidimensional test anxiety scale (Putwain et al., 2020).

The recent measure by Putwain et al. (2020) reflects the change in understanding of the test anxiety construct over the years and moves away from many of the measures that are confined to subscales that measure worry, emotionality, and affective behaviour, in creating a more dynamic and multidimensional scale. The MTAS measures two cognitive dimensions of worry and cognitive interference and two affective physiological interventions of emotionality and autonomic arousal. This new test anxiety measure was not available when I started this study and may have been a good measure to use because although the students were slightly older (11-to 16- years-old) they are from U.K. schools and the measure is multidimensional reflecting what is now known about the test anxiety construct. However, the new measure was created for secondary school aged students and not primary aged students. This situation emphasizes the need for contemporary, appropriate test anxiety measures for the primary age-group.

Putwain et al. (2020) suggest that test anxiety is more than three-dimensional in their construction of a new test anxiety measure, the MTAS or Multidimensional Test Anxiety Scale. My study tries to contribute towards this idea through using an adaptation of the thoughts' subscale from the known Children's Test Anxiety Scale and through the creation of a new measure (Self-efficacy in Test Taking) to capture another dimension of test anxiety.

The principle of using a multidimensional scale to measure a multidimensional construct is one I have tried to support through the exploratory processes I have used with the thoughts' subscale and the adaptations I trialled, and the design of the Self-efficacy in Test Taking questionnaire.

Test anxiety is difficult to measure and is limited to pre and post-test self-report measures. (e.g., Putwain & Pescod, 2018). Zeidner and Matthews (2005) call for improvements in theory and methods of research to reflect the diverse nature of evaluation anxiety. Putwain (2008a) suggests that test anxiety literature is limited by the narrow focus on empirical measurement. Zeidner (2014) refers to a lack of understanding of the individual nature (proximal and distal factors) and argues for the need to crossover between test anxiety and other relevant theories. The Self-efficacy in Test Taking measure that I created may help to fill this gap.

2.2.2.1 Validation for the Use of the Children's Test Anxiety Scale. The main reason that I chose this measure is because it was tested on primary school children and created with the input of a primary school teacher. In addition, previous studies have shown the construct validity of the Children's Test Anxiety Scale (e.g., Aydin & Bulgan, 2017; Putwain et al., 2013; Yeo, Goh & Liem, 2016) The Children's Test Anxiety Scale (Wren & Benson, 2004) was the choice for both the pilot and main study.

Wren and Benson (2004) explain in detail the planning, construction, quantitative evaluation, and validation of the new Children's Test Anxiety Scale. Initially, 230 students were involved from grades 3 - 6 (in England, Year 4 to Year 7). This sample size involves Year 6, which is the year group in this study.

The language used for the items was obtained from the students themselves in a free writing test. The initial 107 items were subjected to a panel of eight teachers from the different year groups and reduced to 50 items. In the experiment, the 50 items scale was delivered to 230 students and resulted in a refined 30 items scale that was again validated using a sample of 261 children, to establish the validity of the construct.

I was able to verify the accessibility of the questionnaire in my classroom as the sample group in the pilot read and answered the Children's Test Anxiety Scale easily. The Children's Test Anxiety Scale is easy to deliver. This aspect is important to consider for the main study. No expert or researcher is needed. A classroom teacher can administer the questionnaire easily and quickly in approximately 10 -15 minutes.

The reliability estimate for the 30 item Children's Test Anxiety Scale was 0.92 and the subscales were 0.85 for *autonomic reactions*, 0.78 for *off task behaviours* and 0.89 for *thoughts*.

I chose The Children's Test Anxiety Scale (Wren & Benson, 2004) because the development, planning, construction, quantitative evaluation, and validation of the questionnaire are detailed by Wren and Benson (2004). The questionnaire is representative of the literature about the test anxiety construct at the time it was developed. The cognitive, behavioral, and affective physiological dimensions of test anxiety are represented through the respective subscales of *thoughts*, *off task behaviours* and *autonomic reactions,* respectively.

2.2.3 Effects of Test Anxiety

This section discusses the relationship of test anxiety and performance and its prevalence in test anxiety literature. Processing efficiency theory (Eysenck & Calvo, 1992) and attentional control theory (Eysenck, Deraksham et al., 2007) are explored in relation to individual performance in tests. The counter theory of test anxiety being a facilitative force is mentioned at the end of this section although this idea has historically and recently been refuted (e.g., Hembree, 1988; Putwain et al., 2020). The concept of students that are high in test anxiety is explored again.

2.2.3.1 Test Anxiety and Performance. This section is brief in comparison to the vast amount of literature that exists about test anxiety and performance. This is because test anxiety and its relationship to performance is not measured in this study. Although the literature about cognitive test anxiety tends to focus on test anxiety and performance (e.g., Hembree, 1988; von der Embse et al., 2018) the actual relationship between test anxiety and performance is small, r = - 0.21 (Eklöf, & Nyroos, 2013; Hembree, 1988; Seipp, 1991). However, the literature about test anxiety and performance is useful to consider in terms of learning how students use cognitive strategies to mediate performance deficits.

Putwain (2008) refers to the relationship of test anxiety and performance, that of r = -.21 and questions why the figure is not larger. He suggests that the answer lies in how students cope with test anxiety. After all, it is not anxiety itself that is responsible for negative performance but the strategies a student uses to cope with and respond to the anxiety.

The literature about test anxiety and performance says that it is the cognitive or worry element of the test anxiety construct that affects cognitive performance more so than the

emotionality aspect of the construct. (e.g., Hembree, 1988). Schwarzer (1990) focusses on the empirical research nature of test anxiety and performance studies and suggests that they are too limited in their scope and lacking in their conclusions.

Test anxiety is not a simple construct to measure because of its complexities and the difficulty of obtaining a measure in highly evaluative circumstances. (e.g., Schutz et al., 2014) Students, who are high in test anxiety, do not always perform poorly (e.g., Putwain & Symes, 2018) and students, who are low in test anxiety, do not always perform well. (Zeidner, 1998)

Students, who have more test anxiety, will often expend more effort, and therefore equalize the achievement outcome, possibly at considerable cost to themselves and their well-being. (e.g., Putwain & Symes, 2018) Conversely, students, who are low in test anxiety, may not have prepared for the tests.

Eysenck, Derakshan, Santos and Calvo (2007) discuss *attentional control* theory, ACT as a development from *processing efficiency* theory (Eysenck & Calvo 1992). ACT interferes with working memory (Owens et al., 2012) and also motivates an individual to reduce threat through increased effort that is typified by the application of more cognitive resources to the high stakes' test. ACT means that performance in exams can be maintained with effort. In *processing efficiency* theory, the effects of test anxiety are said to be greater on processing efficiency rather than *performance effectiveness.* In other words, anxiety can affect the way that information needs to be recalled, affecting working memory, an effect of the worry or cognitive process in the evaluative situation. However, worry or cognitive load doesn't have an impact on the overall result because the brain is able to overcome worry through increased effort and use of auxiliary processing and strategies. In this way, test anxiety is still acting as a debilitating factor, although performance is not affected. The effort expended by the student who is highly test anxious, will be considerable but it is all about how the individual responds to the threat, the test anxiety and whether they have the coping strategies to manage the tense, evaluative situation and what the potential impact is on their well-being.

Putwain and Symes (2018) found that increased effort improved the physiological or bodily symptoms of test anxiety but not the worry element of test anxiety. They suggest that not all students who are high in test anxiety will respond in the adaptive way proposed by the PET *(Pr*ocessing Efficiency Theory) and ACT (Attentional Control Theory) theories as some

will simply find the cognitive load too burdensome or may use avoidance strategies. (Putwain & Symes, 2018)

Owens et al. (2012) suggest that this ability to perform, in students, who have test anxiety, can only happen if individuals have the resources, in terms of working memory capacity, to cope with the negative effects of anxiety. They propose that students, who are low in working memory capacity, simply do not have the resources available to improve performance. Owens et al. (2008) in their study with 10-to 11-year-olds suggest that verbal working memory accounted for 51% of the association between trait anxiety and academic performance. Tse and Pu (2012) support this finding by reporting a negative relationship between students with trait test anxiety and low working memory and use of the testing effect.

However, there are some findings that suggest that with the correct study strategies and / or interventions, that even with low working memory, test anxiety can be reduced through using retrieval practice. Agarwal et al. (2014) suggest that 72% of students felt that retrieval practice made them feel less nervous for end of unit tests. In addition, Agarwal et al. (2017) suggest that the benefits for retrieval practice with feedback are significantly greater $r = -.42$ for students who have low working memory. These findings may challenge Owens et al. (2012) and Tse and Pu (2012).

To summarise, some individuals with test anxiety will not suffer a decrease in their performance or achievements because their anxiety will increase their effort. However, they must work harder than individuals, who have low levels of test anxiety, to achieve the same results. Although, the academic outcome might be the same, because the process demands more from them, it is possible that there is a detrimental effect on the well-being of these individuals. The *facilitating* / *debilitating* issue of test anxiety as proposed by Alpert and Haber (1960) is reduced by these arguments as the effort in processing anxiety is recognized and the issue becomes one of impact on an individual because of the increased effort that is expended.

Finally, Sperling et al. (2017) developed and validated a new test anxiety questionnaire that included the idea of test anxiety being a facilitative influence. They suggest that more students felt test anxiety hindered them than aided performance in tests.

2.2.3.2 Students Who Have High Test Anxiety. As many as 10-20% of students in high stakes' test taking year groups will be high in test anxiety (Putwain & Daly, 2014).

Putwain & Pescod (2018) suggest that there is no universally accepted definition of what constitutes a student who is highly test anxious. Herzer et al. (2014) comment that test anxiety is not coded as a separate diagnosis on the DSM -IV (American Psychiatric Association, 2020) or ICD-10 (World Health Organization, 1992) and suggest that a cut off score of >80 on the Test Anxiety Inventory would denote clinical test anxiety. However, the average age of the students in this study was 25 years old.

According to Hembree (1988), an important aspect about individuals, with high levels of test anxiety, is that they have lower self-esteem than individuals with low levels of test anxiety. Students who are high in test anxiety perceive more self-detrimental than self-enhancing information in the testing process. (Wong, 2008)

Students, who are high in test anxiety, can be characterised by having negative and irrational thoughts, anticipating failure, and experiencing difficulty concentrating in tests and can freeze in testing situations (e.g., Zeidner, 2007, 2014). However, all students are different. In general, students, who are high in test anxiety, lack self-efficacy, belief in their capabilities in test taking for a multitude of reasons, for example: previous failures in tests, trait general anxiety, study deficit skills. (Zeidner, 1998)

Schutz et al. (2002) suggest that, for students, who are high in test anxiety, even learning about test taking strategies, can cause anxiety. Test anxiety is not just found in the exam itself but from the notification of the exam to the performance and then at post-test, in reflections. Zeidner (1998, p. 50-51) refers to these stages as: anticipatory, confrontation, waiting and outcome. Galassi et al. (1981) in a rare experiment where the measures were taken during the actual high stakes' test found that students, who were high in test anxiety, had the most negative thoughts at third point of assessment, in the actual exam. Due to the nature of high stakes' testing, there are real difficulties in obtaining data from actual exams. Future studies may consider collecting data from taking assessments online with artificial intelligence that can read emotions as an answer to obtaining accurate pre and post-test measures for test anxiety (e.g., D'Mello et al., 2014).

Stowell et al. (2008) suggest that for a student, who is high in test anxiety, even completing homework, can be as unpleasant as taking an exam. Bolger (1990) proposes that a student, who is high in test anxiety, is unable to avoid emotion focusing thoughts, at the expense of task related thoughts, during the forethought phase, when the notification of the

exam is given. In turn, their confidence in their study skills is affected as they rate the test more threatening and consequently, prepare less effectively in class.

Sarason (1972) observed that students, who were high in test anxiety were more self-centred and self-critical than students with low test anxiety and experienced more worry. Covington (1984) discussed the intrusive worry that students, who are high in test anxiety experience.

Douglas Wren (2020, personal communication) provided the algorithm for identifying students, who were high or low in test anxiety. In this study, I identified students, who were high in test anxiety. The result was 18% of the sample, $N = 31$. Given the young age of the students and potential benefits of a universal test anxiety intervention (Weems et al., 2015), the intervention was administered to all students in the intervention group.

2.2.4 Summary and Next Section

The transactional nature of test anxiety has been discussed and the need for new measures that reflect the changing dynamic of the test anxiety construct. The idea that students, who are high in test anxiety may use more effort to compensate in tests is explored. This concept may explain why performance deficits due to test anxiety are rather small, r -.21 (e.g., Seipp, 1991). The salient point to note is that it is not the test anxiety that is the issue, but how individuals cope with it (e.g., Putwain & Symes 2018). A further salient point to note is that retrieval practice reduced feelings of test anxiety (Agarwal et al., 2014).

The next section explores the relationships between self-efficacy in test taking, retrieval practice and the concepts of uncertain control (Putwain & Pescod, 2018) and perceived control (Putwain & Symes, 2018) in test anxiety that were all key ideas to the design of the intervention. Chapter 3 explores the intervention itself.

2.3. Self-efficacy

This section is about self-efficacy theory and the idea of control. It is self-efficacy theory, in this case, belief in capability in test taking, that is the theoretical foundation of the additional pre-test and post-test questionnaire, Self-efficacy in Test Taking. This new instrument may provide a crossover measure for test anxiety and other theories (Zeidner, 2014) and an age relevant scale that includes students' perceptions of test situations (von der Embse et al., 2021). The caveat is that the new instrument needs to be cross validated and thus all results are exploratory at this stage.

If an individual is high in uncertain control, then it is likely they will be high in test anxiety (Putwain, 2018). It is through controlling 'control', in this case, through trying to create self-efficacy in test taking feelings, that test anxiety may be moderated.

2.3.1 Theories and Models

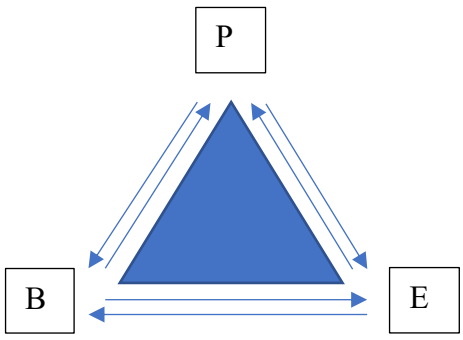
Self-efficacy theory sits under the umbrella of social cognitive theory (Bandura, 1986a, 1986b) and is domain specific about an individual's belief in capability to perform a task. (Bandura, 1997). Bandura (1977, p.192) acknowledges that cognitive processes play a prominent role in relation to behaviour patterns. Moreover, that experiences leave lasting effects in memory. For test anxiety, the strand of self-knowledge beliefs is where access to experience and memory reside to inform beliefs alongside transactions from individual characteristics and the environment.

Social learning theory (Bandura, 1977) posits that we learn from observing each other and social cognitive theory is a development of social learning theory as we learn from corrective feedback on performance so that we learn from ourselves rather than others. The intervention in this study is drawing on self-efficacy theory and therefore social learning and social cognitive theories too.

One of social cognitive theory's central concepts is that of the triad of reciprocal causation. B = behaviour, P = personal factors such as cognitive, affective, and biological events and E = external environment (Bandura, 1986a). All determinants interact to influence each other bi-directionally (Bandura, 1997).

Figure 2.3

Bandura (1986b) Triadic Reciprocal Causation



The elements of this concept are similar to the idea of transactional test anxiety that is dependent upon the dynamic relationship between individuals, their behaviours and the evaluative situation (e.g., Zeidner, 1998).

Bandura (1997) compliments Schunk (1989) and his work that analyses the contribution of efficacy beliefs to levels of cognitive performance. Bandura suggests that "Children with the same level of cognitive skill development differ in their intellectual performances depending on their strength of perceived efficacy". Bandura (1997, p. 216).

In this study, the *perceived efficacy* is a child's belief in their capability to achieve well in a test situation. Intellectual performance is not measured in this study. Improved perceived efficacy may be achieved by delivering an innovative intervention that is aimed at increasing self-efficacy in test taking through knowing about the testing effect.

Sarason (1972) suggested that students who are high in test anxiety are more self-critical and worried more. Bandura (1997) suggests that students who lack belief in themselves need more proximal and frequent feedback of progress and repeated affirmations of their progress. It is feasible that repeating 'your testing routes are well-oiled', the key phrase of the intervention provides such proximal and repeated affirmations of progress.

2.3.1.1 The Social Learning Model in Test Anxiety. The area of social learning theory that focusses on human agency or how an individual influences their thoughts, behaviour and motivation is important in understanding academic self-belief. (Smith et al., 1990). Smith et al., (1990) propose the *social learning* model as one of three models: *cognitive attentional*, *cognitive skills* and the *social learning model*. The *social learning* model in test anxiety (Smith, Arnkoff & Wright, 1990) focusses on a loss of self-control and self-esteem and test taking self-efficacy. They discuss the idea of the development of social learning theory developing into social cognitive theory (Bandura, 1989) as cognitive attentional variables predicted test anxiety and performance in their different models. Smith et al. (1990) suggest there is a need for research into the assessment of negative thoughts and self-efficacy together as the controlling negative thoughts is a powerful "predictor combination" (Smith et al., 1990, p. 319)

The intervention tries to create a belief in capability in test taking or self-efficacy in test taking as described in the social learning model (e.g., Smith et al., 1990) to intervene in the self-knowledge beliefs strand of the S-REF (Self-referent Executive Function) test anxiety model (Zeidner & Matthews, 2005) and manipulate test anxious thinking, at this

point, of the S-REF (Self-referent Executive Function) model (Zeidner& Matthews, 2005) before maladaptive coping can set in. Students, who are high in test anxiety, are more likely to experience personal, negative, worrisome thoughts (e.g., Sarason, 1972) as they access the self-beliefs knowledge strand of the S-REF (Self-referent Executive Function) model (Zeidner & Matthews, 2005) and do not have access to effective coping strategies such as self-efficacy in test taking. (e.g., Smith et al., 1990)

If self-efficacy through knowing about the testing effect, can build resilience, strength to carry on in extreme adversarial school conditions, (e.g., Putwain et al., 2013) or academic buoyancy, everyday school resilience, (Martin & Marsh, 2008) in evaluative situations, anxiety may be alleviated for the *A state* (Spielberger & Vagg, 1995) situation of test anxiety and help individuals who suffer *A trait* (Spielberger & Vagg,1995) anxiety through creating, individual and effective cognitive, self-efficacy practices that can be called upon when needed.

2.3.2 Test Anxiety, Uncertain Control, Academic Buoyancy and Well-being

The idea of *uncertain control* (e.g., Pekrun, 2006; Putwain & Pescod, 2018) refers to the control an individual believes they have in affecting a successful outcome. For the S-REF (Self-referent executive function) model of test anxiety (Zeidner &Matthews, 2005), uncertain control is a self-belief that contributes to the perception of high stakes' tests being threatening (Putwain & Pescod, 2018). Martin and Marsh (2008) state that uncertain control positively correlated with general academic anxiety. Uncertain control means that students cannot understand how their actions are linked to outcomes (Martin, 2002; Putwain & Pescod, 2018). The intervention in this study focusses on one aspect of an individual's control from the S-REF (Self-referent Executive Function) model's (Zeidner & Matthews, 2005) strand of self-knowledge beliefs and that is, self-efficacy in test taking; belief in one's ability to take tests, in the hope of reducing uncertain control by teaching students about the testing effect.

In addition, uncertain control might be reduced if students were allowed some control over the material being tested. Zeidner (1998) suggests that when individuals are allowed some control over computer assisted instruction, state anxiety is lowered. As a teacher, I have used InCAS (Interactive Computerised Assessment Systems) that differentiate the questions according to how well a student is achieving but these were low stakes' tests and did not have senior school entrance or scholarships or G.C.S.E. (General Certificate in Education) grade predictions dependent upon the outcomes.

Teaching students the exam rubric (how papers are set out), taking them to the testing room (if possible) and familiarizing students with how examiners mark are some strategies I have used as a teacher to try and reduce uncertain control. It is possible that teaching students about how the testing effect works, the intervention in this study, may reduce uncertain control.

Putwain and Pescod (2018) propose the idea that there is a need for an increase in recognition that test anxiety is associated with poor student well-being and it is not simply about test anxiety and performance. The nature of the relationship between test anxiety and school well-being is explored further in a more recent paper (Putwain et al., 2021) and the results suggest that school-related well-being is negatively related to test anxiety but not the other way around.

Nevertheless, Putwain et al. (2021) conclude that there would be downstream benefits to better mental health if test anxiety interventions are used. Steinmayr et al. (2016) state that 33rd percentile of test anxiety scores can reliably predict clinical anxiety and that the worry element of test anxiety negatively predicted life satisfaction. Weems et al. (2009) suggest that test anxiety interventions in schools can reduce general anxiety and thereby contribute effectively to increasing students' overall well-being.

Test anxiety poses a risk to well-being, particularly vulnerable students (Putwain et al., 2014). These are the students, who are low in academic buoyancy (Putwain et al., 2014). Recently, Putwain et al. (2020) demonstrated that test anxiety is positively correlated to an increase in mental health risk and lowers well-being in school. However, Putwain et al. (2021) suggest that school-related well-being is negatively related to test anxiety but not the other way around.

These findings illustrate the potential impact of test anxiety to a student's welfare. If left untreated, it may reduce the well-being of students in school (Putwain & Symes, 2018). The number of students who are test anxious is worrying and is still not being taken seriously enough (Putwain, March 2021, personal communication).

Steinmayr et al., (2016) discuss the idea that the relationship between test anxiety and well-being is not the focus of many studies and that the literature tends to be focused upon the test anxiety and academic performance.

A potential solution to the problem could be found in Weems et al. (2009), who suggest using a test anxiety intervention that is administered universally, as was the case in

this study, because that may increase general student well-being. On a personal note, I suggest another answer may lie in curricula reform and in the way that students are tested (see Chapter 10, Reflections)

Many of Putwain's studies are with Key Stage 4 students (Year 10 and 11 and students are 15-to16- years - old.) The focus is on reducing test anxiety around the high stakes' G.C.S.E. (General Certificate of Secondary Education) tests. Putwain et al. (2013) identified resilience as being linked to test anxiety and test performance. The authors call for resilience programmes in schools.

A resilience programme is found in Putwain and Pescod (2018) and the 'STEPs' (Strategies to tackle exam pressure and stress) intervention for Key Stage 4 (15-to16 year-olds) students; a successful intervention that combines relaxation, study skills training and replacing negative with positive self-talk. Putwain and Pescod (2018) appeal for interventions that will reduce anxiety and fear of failure. This call is described as a "pressing need" for Year 6 (the ages of the students in this study) in a paper about examination pressures on children and young people (Putwain, 2020, p.6).

Putwain, Daly et al. (2016) discuss the concept of academic buoyancy and its impact on test anxious students. They acknowledge the study's limitations as it largely (like most test anxiety research) relies on self-report measures, from which there is a danger of self-presentation bias. They conclude that training students in academic buoyancy could improve the cognitive or worry element in test anxiety. Being self-efficacious in a domain, or aspect of school life could contribute towards feelings of academic buoyancy.

Academic buoyancy is a concept distinct to academic resilience (e.g., Putwain, Daly et al., 2016; Martin, 2013). The principle is that academic resilience is the characteristic displayed when the academic or school situation demands an extreme form of strength. Academic buoyancy is an everyday resilience that is needed by students in schools.

Putwain, Daly et al. (2015) examined whether students' academic buoyancy affected the negative relationship between cognitive test anxiety and performance in high stakes' exams. The 5Cs' of academic buoyancy are listed as: confidence (self-efficacy), coordination (planning), control (persistence), composure (low anxiety), and commitment (persistence). They found that worry predicted a lower G.C.S.E. (General Certificate of Secondary Education) mean score and academic buoyancy a higher G.C.S.E. (General Certificate of Secondary Education) mean score.

To summarise, there is a worrying number of students who are test anxious in schools and this state is likely to be linked to their overall well-being. Knowing that test anxiety likely impacts upon student well-being means that the focus should be on controlling uncertain control and managing perceived control through employing test anxiety interventions (and curricula) that promote academic buoyancy and resilience and in creating interventions for Year 6 where there is a "pressing need" (Putwain, 2020, p.6).

2.3.2.1 Coping Strategies. Putwain (2008) discusses the effect of test anxiety on performance and questioned why the relationship is not larger (r = - .21). The answer lies in how individuals cope.

Students, who experience test anxiety, need coping strategies. The trait and state forms of test anxiety (Spielberger and Vagg, 1995) mean that the distal antecedent of *A trait* anxiety can combine with proximal antecedent of *A state* to result in an individual, who is high in test anxiety. In addition, individuals who are test anxious, are all different. All will cope in different ways and some more successfully than others. Students who have general (trait) anxiety that manifests as state anxiety in evaluative situations may have more coping strategies than someone who suffers with state test anxiety and who does not know how to cope with high stakes' exams. Some examples of negative coping strategies: rumination, avoidance, and self-blame (Putwain, 2018).

One coping process that may alleviate test anxiety may be found in retrieval practice. The intervention in this study uses retrieval practice to try and reduce test anxiety but deviates from the use of retrieval practice to secure academic, subject content to recall for high stakes' tests and uses retrieval practice to recall the testing effect as a strategy to create self-efficacy in test taking for high stakes' tests.

Agarwal et al. (2014) report that 72% of middle to high school students felt that retrieval practice made them less nervous for tests and exams. Although 19% reported feeling greater anxiety as a result of using clicker quizzes. It is possible that retrieval practice about the testing effect alongside practice tests may also make students feel less nervous about exams. (Barsham & Ellefson, 2020). To my knowledge, these two papers represent the extent of the extant literature about using retrieval practice to reduce test anxiety with 10-to11-year-olds and Agarwal et al., (2014) is with 11-to 16-year-olds so only a percentage of the sample is comparable to the U.K. Year 6 (10-to 11-year-olds).

Schutz et al. (2008) describe the process of regulating emotions related to testing as a fluctuating state that can range from pleasant to unpleasant. They say what is important is how students regulate their emotions at these times. They suggest that the emotion of anger can protect the test anxious individual. They add that if we try to understand how students can control their emotions in tense, evaluative situations then it is easier to help them. The need for an individual to feel in control is discussed and a need to keep the exam itself in perspective. To view the exam as important but not overly so, in comparison to other aspects of life. Schutz et al. (2008) explore the idea of agency that is the extent to which students judge themselves as being in control of the evaluative situation. The idea of reducing uncertain control (Putwain & Pescod, 2018) and improving perceived control (Putwain & Aveyard, 2018) are important ones to the design of the intervention in this study that tried to give students more feelings of control about their test taking abilities.

As per the S-REF (Self-referent Executive Function) model (Zeidner & Matthews, 2005), an individual will call on their self-knowledge beliefs and make a judgement about how suited they are to control the evaluative situation. What follows next is either adaptive or maladaptive coping strategies. The judgements that students make about their ability to handle the test can emerge before, during and after tests. Test anxiety is likely to occur when students feel they have no control. (Zeidner, 1998) Therefore, one way of coping may be through receiving a test anxiety intervention that targets control in the self-belief knowledge strand of the S-REF (Self-referent executive function) model of test anxiety (Zeidner & Matthews, 2005).

Zeidner and Matthews (2005) refer to performance deficit theories that focus on personal characteristics that can lead to elevated and reduced levels of test anxiety. The reality, for some individuals, is a miserable mix of negative thoughts in the self-referent state causing low self-esteem and often an underestimation of the self. For an individual with low self-esteem, coping occurs through rumination or worry, self-depreciations, self-handicapping as a means of maintaining self-worth, for example not bothering to study and extolling this action as an attribute, whereas in reality, it can be masking ability or test anxiety.

Matthews, Hillyard and Campbell (1999) used the S-REF (Self-referent Executive Function) model (Zeidner & Matthews, 2005) of test anxiety to look at coping strategies from long-term memory. They found that short-term intense distress and worry are generated by accessing negative self-beliefs, the middle strand of the S-REF (Self-referent Executive

Function) model. It is at this point that individuals may choose unproductive coping strategies like self-blame and employ avoidance metacognitive beliefs (e.g., if I worry about the exam, it will be all right) that maintain negative self-referent thinking. This strand of the S-REF (Self-referent Executive Function) model (Zeidner & Matthews, 2005) has been targeted by the intervention in this study so that students can access positive self-beliefs rather than negative ones from this place of the test anxiety construct.

The role of a teacher and the relationship with a student (e.g., Hattie & Yates, 2014) may be of key importance in managing test anxiety (e.g., Schutz et al, 2014). Using fear appeals *appropriately* may be a coping strategy. Usually, teachers are held to account for their students' performances in high stakes' exams and can become tense themselves as the exam approaches. Putwain et al. (2016) urge for these claims about exams to be conveyed with caution. For students who are low in test anxiety, these fear appeals have shown that they encourage the students to make more of an effort but for students, who are high in test anxiety, they can make the test anxiety worse. (Putwain et al., 2016). Putwain's call for attributional training for teachers may contribute towards helping students cope in evaluative situations. (Putwain & Aveyard, 2018)

Finally, writing about test anxiety can help students to cope with evaluative situations (e.g., Lang and Lang, 2010; McKeachie, 1951; Ramirez & Beilock, 2011). This idea is important to this study because the intervention has two writing tasks in it. This idea is explored further in Chapter 3.

2.3.2.2 Perceived Control and Self-Efficacy in Taking Tests. Self-efficacy theory (Bandura, 1997) is about an individual's perception of their capability in managing a resulting situation. Zeidner and Matthews (2005) state that self-efficacy is more strongly related to successful performance than is self-esteem. Correlations of -.30 to -.66 are recorded of the relationship between self-efficacy and anxiety by Schwarzer & Jerusalem (1992). For Smith et al. (1990), self-efficacy for test success contributed to the prediction of test anxiety ahead of cognitive interference and poor study skills.

Putwain and Aveyard (2018) discuss the issue of self-beliefs showing a small to moderate correlation with exam performance. They state that there are no studies, at the point at which the paper was written, that have examined how self-beliefs might interact with cognitive test anxiety, to predict exam performance. In this study, the self-belief of self-

efficacy in test taking has been measured to see how it relates to the worry element or cognitive test anxiety, but not in relation to performance.

Perceived control is how certain a student feels about their ability to perform in an evaluative situation (e.g., Putwain & Aveyard, 2018). Perceived control is different to uncertain control because a student can see and understand the link between what they do, how they approach a test and the outcome (Putwain & Aveyard, 2018). Therefore, it is perceived control and uncertain control that the intervention in this study tries to manipulate to afford a student more control over their test taking feelings.

Putwain and Aveyard (2018) propose that perceived control, academic buoyancy, and test competence moderate the strength between worry, or cognitive test anxiety and exam performance. They examine the relationship between worry and perceived control. As worry, or cognitive test anxiety increased, higher perceived control decreased.

In addition, younger children and girls reported less perceived control. Older children have more perceived control than younger children. The regulative effect of control on anxiety may be less strong in younger children. (Putwain & Aveyard, 2018). My study may be able to add to these ideas about perceived control and worry for this younger age-group because the intervention aimed to increase feelings of self-efficacy in test taking to reduce worry about taking tests in 10-to 11-year-olds.

Smith (1989) discusses the desirability of helping people develop a sense of self-efficacy because self-efficacy is highly correlated to achievement. Self-esteem and self-efficacy are consistently found to have the strongest relationship with test anxiety (Von der Embse et al., 2018) but the authors note that correlation is not causation.

It is self-belief that is critical to a students' motivation and resilience. Self-efficacy and self-belief are in turn linked to self-regulation, effort, persistence, resilience, academic buoyancy, and subsequent achievement. (Bandura, 1997)

Schunk and Pajares (2005) discuss the idea of competence perceptions and perceived control and propose that people expect success relative to the measure that they feel in control of their behaviour. If an individual has a sense of control over the desired outcomes of their life, then this control is a key motivator for their behaviour. Moreover, self-efficacy beliefs are sensitive to contextual factors and are different to other self-beliefs because they are more task and situation specific. Self-efficacy is domain specific (Bandura, 1997). In this study, the intervention was designed to try and incorporate the ideas of self-efficacy in test

taking being task and situation specific. Teaching students specifically about how the testing effect works to allow them to apply this knowledge to evaluative situations tried to give them mastery over the domain of self-efficacy in test taking.

Self-efficacy is generally assessed on a microanalytic level (Schunk & Pajares, 2005). This level appears as asking individuals to report the level, generality, and strength of their confidence to accomplish a particular task. The design of the questionnaire in this study was designed to assess levels of Self-efficacy in Test Taking and tries to follow this microanalytic idea but the new measure needs evaluating on another sample.

2.3.3 Creating a Self-efficacy Measure; the Self-efficacy in Test Taking Instrument

From the pilot, I was able to conclude that I needed a measure for self-efficacy in taking tests at the pre-test and post-test stages to see if the cognitive intervention affects feelings of self-efficacy. An original feature of this study is the creation of an instrument which measures self-efficacy in test taking in 10-to 11-year-olds.

I designed the measure using Bandura's 'Guide for Constructing Self-efficacy Scales' (Bandura, 2006) and followed the principles outlined below in the construction of this measure. The italics represent the thinking behind the examples detailed.

1. Clear definitions of the activity in the domain of interest. For example, 'If I find a test hard, I keep trying'. *Questions focussed on the test taking situation.*

2. Good conceptual analysis of its different facets. For example, 'I can concentrate in tests even if I find them difficult'. *Acknowledging the worry aspect and potential off- task behaviours, cognitive interference and maladaptive coping which may occur.*

3. Good conceptual analysis of the types of capabilities it calls upon and the range of situations in which the capabilities might be applied. For example, 'If I find a test hard, I feel I will not pass'. *Almost the opposite of the example given above in 1 -representing the different extremes found in individuals who take tests.*

4. Efficacy beliefs differ in level. A four- point Likert type scale was used. Indeed, the questionnaire used a very similar Likert type scale as the Children's Test Anxiety Scale measure. *This decision was deliberate and aimed to 'keep it simple' for the 10-to 11-year-olds in this study. I did not use a 10-point scale as recommended by Bandura in the guidance as I felt it was difficult for this age-group.*

5. I also tried to avoid generality, indefiniteness, ambiguity, and omnibus measures.

I did not use the ten- point Likert scale recommendation as my experience as a primary school aged teacher suggested a ten-point scale would give the students too many options to choose from and may cause some confusion. Instead, I used a similar four-point Likert type scale as found in the Children's Test Anxiety Scale (Wren & Benson, 2004). I used my knowledge as a teacher of primary children and drew upon the many conversations I have had with students of this age-group about tests to make an accessible and easily interpretable questionnaire. In addition, I presented the questionnaire to the INSTRUCT (Implementing New Student Thinking Resources Using Cognitive Theory) lab to consult some expertise and received feedback about reversing the wording of some of the questions.

Furthermore, I tried to represent the situational aspect of social cognitive theory, the transactional nature of test anxiety by including questions about the instructor and the place or classroom where the tests were taking place. (e.g., von der Embse, 2021) However, these two questions were the ones that I removed from the questionnaire for the analyses as the skew and kurtosis were not at acceptable levels and the exploratory and confirmatory factor analyses suggested it would be better to remove these items. (see Chapter 7). Finally, I trialled the questionnaire on my Year 6 class before other teachers were given it to use.

2.3.4 Summary and Next Section

This section of this chapter has explored how an aspect of social cognitive theory, self-efficacy theory and its relationship to academic buoyancy and resilience is potentially a key influence in being able to manipulate an individual's perceived control and reduce uncertain control, as they reside in the self-knowledge beliefs strand of the S-REF (self-referent executive function) model of test anxiety, to reduce worry or the cognitive element of test anxiety in evaluative situations. If students had more control over their tests, it is likely that state test anxiety would reduce (Zeidner, 1998).

I have considered some of the vast amount of literature involved in three theoretical domains: the cognitive theory of: the testing effect (and desirable difficulties), test anxiety and self-efficacy. The emerging principles for my intervention are that using retrieval practice of the testing effect itself may give confidence from knowing about the testing effect, "your testing routes are well-oiled" to enable self-efficacy in test taking to manipulate the middle strand of the transactional S-REF (Self-referent executive function) test anxiety model (Zeidner & Matthews, 2005), self-knowledge beliefs, to try and stop test anxiety before it starts.

In the next chapter, I briefly explain how my intervention has fulfilled the criteria for an intervention (Fraser et al., 2009), examine the literature around test anxiety interventions and with particular reference to the age-group of this study. The intention is to provide context for the reader about the nature and design of the intervention in this study. Although, as the design of my intervention has never been tried before with any age-group, there is nothing specific in the literature that I can refer to. However, I have examined the test anxiety literature for recommendations for future interventions and hope that the reader is able to observe how the intervention in this study is contributing to the salient suggestions of several authors.

3. Intervention

The intervention that was designed to address the problem of cognitive test anxiety was delivered in primary classrooms and used a combination of three theories as the basis for its design: the cognitive theory of the testing effect (e.g., Roediger & Karpicke, 2006b), test anxiety theory (e.g., Zeidner, 1998) and self-efficacy theory (Bandura, 1997). In this way, the intervention contributes to the advancement of theoretical understanding that may be able to help the work of others (McKenney & Reeves, 2013) in the field of trying to reduce test anxiety in primary aged students through using retrieval practice. (Agarwal et al., 2014) The intervention was delivered by primary school teachers and me (as a teacher and researcher) in primary classroom settings and is underpinned by theory; such interventions are scarce in the literature and in practice (e.g., Putwain, 2020)

One further idea behind the design of the intervention is outlined by Yeager and Walton (2011). The paper is a meta-analysis of social-psychological interventions in education and their impact on achievement. They suggest that these types of interventions do not have an academic agenda but instead focus upon students' psychology. The intervention in this study may fit into the social-psychological category. Achievement is not measured in this study. Yeager and Walton (2011, p.267) caution the reader against finding social-psychological interventions as 'magical' but acknowledge their success as they are able to focus on students' personal feelings and experiences in school in subtle and persuasive ways whilst simultaneously conveying psychological ideas. "Well-oiled testing routes" was the phrase used in this intervention and may be a persuasive, social-psychological technique.

The intervention follows the steps suggested by Fraser et al. (2009, p.36) although not all of these steps are fully completed at this stage as further refinement may be appropriate (see Chapter 10 for the evaluation of the intervention): 1. Specify the problem and develop a program theory. The problem is test anxiety, and the program theory is a combination of self-efficacy, test anxiety and the testing effect theories; 2. Create and revise program material (see Appendix A for Questionnaires, Chapter 7 for preliminary analyses of measures and Appendix D for the training manual); 3. Refine and confirm program components. This refinement was largely done through factor analyses and through consulting the INSTRUCT lab (Implementing New Student and Thinking Resources Using Cognitive Theory). The evaluation of the intervention in Chapter 10 has further suggestions about this step based on teacher feedback about the intervention; 4. Assess effectiveness in a variety of settings and

circumstances. There was a pilot and lessons were learned from the pilot so that another, new questionnaire was created, and a control group used for the main study, see also the Results in Chapter 8 and Discussion in Chapter 9; 5. Disseminate findings and program materials: (see page ii)

This chapter provides an overview of the literature about test anxiety interventions. In addition, aspects of design-based theory that were used to create the intervention are also explored. The fidelity and validity of my intervention are considered.

3.1 Test Anxiety and Interventions

There is a lack of research about test anxiety interventions with 10-to11- year-olds. (e.g., Ergene, 2003; Lohbeck, Nitkowski & Petermann, 2016; von der Embse et al., 2013; Yeo, Goh & Liem, 2016). There is a need for test anxiety interventions with school aged populations (e.g., Putwain & Symes, 2018). Putwain (2020, p.6) refers to a "pressing need" for a test anxiety intervention for the primary year group, Year 6 (10-to 11- year-olds).

Interventions that address the behavioral aspect of test anxiety, like desensitization (Ergene, 2003; Hembree,1988; Spielberger et al., 1976; von der Embse et al., 2013) lessen test anxiety but do not improve performance. Self-esteem, self-efficacy, and cognitive control are key constructs to bear in mind, when designing an intervention. (e.g., Putwain & Pescod, 2018; Weems et al., 2015). It is a combination of behavioural interventions with cognitive intervention and study skills that can improve performance (e.g., Spielberger et al., 1976).

Test anxiety interventions tend to be rooted in clinical psychology and use forms of cognitive behavioural therapy. I think it is worth differentiating what I mean when I refer to my intervention as a cognitive intervention. I am not referencing a clinical psychology background where a combination of cognition and behaviour therapies are used to treat individuals but cognitive psychology where there is less emphasis on an individual and where cognitive mechanisms are being examined across the board. In this study, I am looking at how our neural architecture (personal communication with Dr. Michelle Ellefson, March 2021) can be and has been altered by an intervention that uses cognitive theory. It may be that by manipulating thinking about self-efficacy in test taking that feelings about self-efficacy in test taking can be altered.

Putwain and Aveyard (2018) focus on how three self- beliefs: *academic buoyancy*, *perceived control* and *test competence* affect the relationship between worry and test performance. The authors conclude that self-belief might alter the problem of cognitive test

anxiety and examination performance. Putwain and Aveyard (2018) suggest interventions that will reduce worry, the cognitive element of test anxiety, to improve an individual's belief in their self-control are beneficial; one's ability to manage the evaluative situation.

Putwain and Pescod (2018) reflect upon the lack of literature for school-based interventions in test anxiety, with the exceptions of: Ergene (2003); Von der Embse et al. (2013); Weems et al. (2015) and Yeo et al. (2016). Thus, this school-based intervention, aimed at 10-to11- year-olds may add to the literature about test anxiety school-based interventions for this age-group.

von der Embse et al. (2013) conducted an analysis of multiple studies to analyse the effectiveness of interventions and identified that a combined approach of both behavioural and cognitive aspects in test anxiety interventions are more effective. However, this analysis was conducted on only ten studies. The authors could only find ten studies that evaluated the effectiveness of a test anxiety intervention.

Weems et al. (2015) examine the longitudinal effects of an intervention in test anxiety. They suggest measuring interventions to see if the positive effects that are recorded initially, from a test anxiety intervention are maintained. They identify the need for longitudinal studies about the effectiveness of test anxiety interventions. They delivered the intervention as a universal intervention. The principle is that all students can benefit from support. A universal intervention means that individuals are not at any risk of becoming the focus of unwarranted attention from peers. The intervention in this study was delivered universally.

In addition, Weems et al. (2015) support the idea of the importance of intervening in test anxiety, by applying a universal intervention to improve overall well-being. The authors suggest that by lessening the anxiety around evaluative situations, through effective interventions, students will achieve more and thus feel more confident about their results and thus become more self-efficacious.

3.1.1 Intervening in Cognitive Test Anxiety

Agarwal et al. (2014) state that using retrieval practice with middle and high school students made 72% of them feel less anxious about end of unit tests and exams. Retrieval practice itself as an intervention. In my study, retrieval practice of retrieval practice (the testing effect) was used to try and create feelings of self-efficacy in test taking and because retrieval practice has been shown to improve recall for students with low working memory (Agarwal et al., 2017), as a teacher, the findings from Agarwal et al. (2014) suggest that my intervention would be inclusive for these students.

Interventions are effective in reducing test anxiety but there is a lack of research into the effectiveness of interventions for 10-to 11- year-olds. (Putwain, 2020). There is a lack of research into the effectiveness of test anxiety programmes for school aged students. (e.g., Ergene, 2003). The evidence base for school-based intervention programmes is extremely limited. School-based interventions have the potential to lessen or even prevent educational underachievement and the threat to well-being that can be the result for students, who are high in test anxiety, as a result of possessing low academic buoyancy, low self-efficacy. (Weems et al., 2010, 2015)

Young et al. (2003) call for the development of intervention approaches based upon modification of schema; moderating how students, who are high in test anxiety think.

Putwain and Aveyard (2018) suggest that interventions may not necessarily improve exam performance unless they also improve *control*. Interventions need to focus on reducing the *worry* or cognitive element of test anxiety to increase *perceived control*. Putwain and Aveyard (2018) propose that teachers, educators should be trained in attributional training so that they can manage student responses to the causes of success or failure. The authors suggest that teachers could play a pivotal role in the management of test anxiety. It is my opinion that the relationship between a teacher and student (e.g., Hattie & Yates, 2014) is one that could mitigate against test anxiety in schools.

3.1.1.1. Individual Interventions Pintado and LLamazares (2016) suggest creating individual interventions to address the individual needs of students' levels of test anxiety. As a teacher, I support this idea but wonder about the difficulty in implementing individual interventions in a busy, classroom environment and also in the respect of fulfilling stringent exam board regulations that are in place to ensure equality of opportunity.

Birenbaum and Nasser (1994) state that individuals with high and low achievement abilities need to be treated differently in terms of intervention. Students, who are highly anxious and high achieving need test taking skills, whereas students, who are highly anxious, and low achieving students would benefit more from training in effective learning strategies. Both groups benefit from cognitive coping techniques, but they need to be different; differentiated training in test taking skills.

There is evidence of students, who are high in test anxiety, performing better under relaxed conditions and students, who are low in test anxiety, benefiting from conditions that cause some test anxiety (Wong, 2008). As a teacher, I think that the reality of being able to provide this environment in the real life setting of a classroom or exam room for a high stakes' test could be difficult to achieve. Whilst the idea would likely benefit the students, the idea of being able to differentiate high stakes' test instructions goes against the equality of provision that is needed for administering an exam.

However, some students, with individual needs do receive different support and access arrangements for exams. For example, extra time can be awarded or the support of a reader or scribe. Some students with general and test anxiety do have different access arrangements for high stakes' tests. A potential problem is that for some students test anxiety is a state (Spielberger & Vagg, 1995) that happens as a result of the knowledge of the test and thus may not be diagnosed and catered for ahead of the test itself. This is one reason why, in my opinion, universal interventions are a good way to manage test anxiety. Alternatively, an effective means of diagnosing test anxiety is needed (Putwain et al., 2020; von der Embse et al., 2020) so that appropriate arrangements can be made for taking high stakes' tests.

3.1.2 Writing About Test Anxiety

Ramirez and Beilock (2001) found that writing about testing worries boosts exam performance in the classroom. A similar idea was discussed by Spielberger and Vagg (1995) with reference to the work of McKeachie (1951). McKeachie (1951) found that anxious students achieved more on multiple choice questions, taken in exams, if they could write comments about the questions. McKeachie (1951) described the opportunity as being able to lose or direct some of the tensions created by tests.

Yeager et al. (2013) use the idea of the brain's plasticity and malleability to promote Growth Mindset about issues such as bullying and aggression in 14-to 16- year-olds. In this

study, the students complete a writing assignment that emphasizes using positive coping skills and avoiding negative coping skills.

Lang and Lang (2010) suggest that students, who wrote down thoughts about being competent or self-efficacious in the situation, lessened the association between test anxiety and test performance by increasing the performance of test anxious students. However, for students, who were low in test anxiety, the same activity lowered their performance.

Writing about the testing effect in a 'Letter to Bob' to allay Bob's fears about the way students learn and writing a letter to a friend to give advice when they are worried about taking tests are writing activities in the intervention in my study.

3.1.3 Controlling Control for Self-efficacy in Test Taking and Well-being

Putwain and Pescod (2018) suggest that *uncertain control* is a plausible mediator in the test anxiety construct and the analysis of mediators can help the researcher understand why and how interventions work. It is these states of *uncertain control* (Putwain & Pescod, 2018) and *perceived control* (Putwain & Aveyard, 2018) that the intervention in this study was designed to mediate by providing more control over the self-belief of self-efficacy in test taking.

The intervention was designed to teach students about the testing effect to give them the 'why' testing works. Giving students the 'why' gives them more control and students who have high control should have access to more positive self-beliefs and therefore lower worry or cognitive test anxiety (Putwain, 2018; Putwain & Aveyard, 2018).

Putwain, Daly et al. (2016) discuss the relationship between students' self-report levels of test anxiety and academic buoyancy and suggest that providing in school training in task focus and orientation would help with coping processes. The authors suggest training students in the 5Cs of academic buoyancy: confidence, coordination, control, composure, and commitment. They provide a mechanism: through individualising work, setting goals, using feedback to reinforce the link between effort and academic outcomes and suggest that this mechanism will help students to manage academic pressures and may help to lessen test anxiety.

Martin et al. (2010) conducted an experiment about the longitudinal modelling of academic buoyancy. They conclude that interventions designed for students' academic buoyancy should target students' self-efficacy. Self-efficacy in test taking is the target of the intervention in my study.

Steinmeyr et al. (2016) suggest that test anxiety interventions that focus on the worry element of test anxiety may have a positive effect on a students' subjective well-being. They suggest that strategies for overcoming test anxiety need to focus on reducing the risk factors for developing test anxiety in the first instance. The design of the intervention in my study aimed to reduce risk factors by using retrieval practice in the lead up to high stakes' tests to teach the testing effect to encourage self-efficacy in test taking.

Weems et al. (2010) call for interventions that can benefit the well-being of the whole school population and feel that an intervention designed to reduce test anxiety may almost inadvertently contribute to overall student well-being in school.

Recently, Putwain, Gallard et al. (2021) found that school-related well-being is negatively related to subsequent test anxiety but not vice versa. They suggest that more studies are needed to explore the relationship between test anxiety and students' well-being.

3.1.4 The Design of My Intervention

The intervention in this study used retrieval practice of the testing effect to try and improve self-efficacy in test taking to reduce worry. The design of the intervention seeks to manipulate the self- knowledge beliefs strand of the S-REF (self-referent executive function) model of test anxiety (Zeidner & Matthews, 2005) to alter an individual's perceived control (Putwain & Symes, 2018) and reduce uncertain control (Putwain & Pescod, 2018) about taking tests. The focus was on worry only, the cognitive element of test anxiety and thus only cognitive strategies were used in the intervention to try and alter feelings and behaviours about test taking.

The literature review suggested that writing about tests can help with test anxiety (Lang & Lang, 2010; McKeachie, 1951; Ramirez & Beilock, 2001). Two weeks of the six-week intervention in this study are devoted to writing. Firstly, there is the letter to Bob Bjork explaining what has been learned about the testing effect. The idea of recalling and writing this knowledge is effectively practising retrieval practice about retrieval practice.

A second task involves writing a letter to a peer who has said they are feeling anxious about a test. The idea is to use the strong, peer associations which can form at around this age to promote positive thinking about test taking and to use the knowledge gained from the cognitive intervention of the testing effect to reassure an anxious friend. For examples of the writing tasks (see Appendix D)

The idea of improving well-being, from an intervention, for the whole school population (e.g., Weems et al., 2010) is important in this study but well-being is not measured specifically. Steinmeyr et al. (2016) suggest that a focus on the risk factors to avoid the onset of test anxiety is the best solution for improved student well-being. The design of the intervention in this study aimed to avoid debilitating test anxiety by using retrieval practice in the lead up to high stakes' tests to teach the testing effect to encourage self-efficacy in test taking. In addition, the intervention was delivered universally. It was practical and inclusive and may have improved students' general well-being (e.g., Weems et al., 2010).

The intervention in this study is not curriculum subject specific but can be used in any lesson and at any point, prior to high stakes' testing. It is easy to deliver and designed by a classroom teacher.

3.2 Design-Based Theory - Intervention

Anderson and Shattuck (2012) suggest that the design of an intervention is key to the quality of results. Aspects of design-based research are a model for this study (e.g., Anderson & Shattuck, 2012; Cobb et al., 2003; McKenney & Reeves, 2013). Aspects of design-based research in this study include: 1. The research was situated in a real educational context, the messy setting of primary classrooms (e.g., McKenney & Reeves, 2013); 2. The study includes the design and testing of an intervention to address a problem in practice, a real world intervention (e.g., McKenney & Reeves, 2013); 3. The intervention could be easily administered in any classroom by any adult (e.g., Ann Brown, 1992); 4. There is an empirical investigation yielding theoretical understanding, theory building and practical applications that can inform the work of others (e.g., McKenney & Reeves, 2013). I have tried to be pragmatic and theoretical in the design of my intervention. (e.g., Cobb et al., 2003) and these aspects I have listed are not exhaustive of the literature.

These principles of the design-based experiment are added to by Cohen et al. (2017) with additional emphasis on the pedagogical goal to be investigated and reference to the factors that may inhibit or enhance the effectiveness. They also suggest that data collection uses multiple methods and is iterative and that the intervention is adapted to appeal to all stakeholders. In addition, Cobb et al. (2003, p.9) refer to the "learning ecology" of a design experiment and the "crucibles for the generation and testing of theory". The extent to which I have been able to manage these principles is evaluated in Chapter 10.

McKenney and Reeves (2013) suggest that the intervention needs to be designed and tested using a mixed methods approach and with multiple iterations. In this study, there is some feedback (see Appendix D) to ascertain the effectiveness of the intervention in the form of teacher and pupil questionnaires that were completed after the delivery of the intervention. The implementation of the intervention in this study is reviewed in Chapter 10.

The teacher and pupil questionnaires that ask for feedback from the intervention reinforce the idea of the researcher / practitioner relationship and supports the idea of establishing fidelity of the intervention (e.g., Cobb et al., 2003; Murphy & Gutman, 2012).

The design of this study is different to action research because of its emphasis on theory to create the intervention (Cobb et al., 2003). The design of the intervention, in this study is the combination of three theories: the cognitive theory of the testing effect (e.g., Roediger & Karpicke, 2006b), test anxiety theory (e.g., Zeidner,1998) and self-efficacy theory (Bandura, 1997). It is the combination of these three theories that do real work (Cobb et al., 2003, p.10).

The experiment is designed to address the problem of test anxiety that is both scientifically and practically important in school evaluative situations. This study demonstrates how the cognitive theory of the desirable difficulty, the testing effect, (Bjork, 1975) can be used in a primary classroom to try and increase feelings of self-efficacy and potentially reduce test anxiety in 10-to 11-year-olds.

Table 3.1 summarises how the intervention in this study meets the elements of design-based theory.

Table 3.1

Meeting Design-Based Theory (Mckenney & Reeves, 2013)

Design Based Theory	This Study
1. The creation of an intervention to promote learning goals in the primary classroom.	1. The learning goals are learning how the testing effect works and gaining self-efficacy in test taking from this knowledge (Bjork, 1975).
2. The development of a theory to explain how the intervention supported the learning goal; advancement of theoretical agenda (Barab & Squire, 2004)	2. The theory that is being developed is that teaching primary aged students about the testing effect (Bjork, 1975) increases feelings of self-efficacy and reduces the cognitive or worry element of the test anxiety construct. Using the combination of self-efficacy theory, test anxiety theory and the testing effect theory.
3. Is the intervention something that would not be seen in the typical primary classroom?	3. It is not common practice in schools to teach cognition of the testing effect (Bjork, 1975) to help students cope with evaluative situations.
4. The concept of the messy setting; actual primary classrooms.	4. Actual, messy primary school classrooms. (Barab & Squire, 2004)

3.2.1 Fidelity of Implementation

Fidelity of implementation is important. It means the extent to which teachers in this study implemented the intervention, as I intended (e.g., Dusenbury et al., 2003; McIntyre et al., 2007). Dusenbury et al. (2003) refer to a broad set of fidelity definitions. Furthermore, fidelity of implementation may help in the achievement of improved outcomes and give confidence in attributing the outcomes of the intervention. (Caroll et al., 2007). Table 3.2 represents the requirements and the steps I took to fulfil them.

Table 3.2

Fidelity of Implementation, Dusenbury et al., (2003)

Fidelity Measure	This Study
Adherence to the programme	Teacher's Manual / Teacher Questionnaire
The amount of programme delivered	Teacher Question 'Which Cognitive Worksheets Did You Use?'
Quality of programme delivery	Opinions invited for improvements and specific questions relating to the worksheets.
Participant responsiveness	Questions about the students' participation and understanding.
Programme differentiation	First three questions on the teacher questionnaire focus on *the testing effect.*

Increasing treatment fidelity is an important consideration. Hulleman and Cordray (2009) examined the difference in effectiveness between field and laboratory experiments concluding that laboratory experiments are more effective. One of the main reasons is attributed to the teachers in a classroom. For example, if the students are not given the correct materials, then student responses could be skewed.

In this study, one school did not include the student responses about the intervention, but I have controlled for this eventuality as both the student and the teacher questionnaires elicit the same information. For example, students were asked 'Were you able to use any of the strategies we have been discussing?' and teachers were asked 'Have the children

understood the testing effect? Furthermore, I was able to deliver the intervention myself to two classes.

I designed the intervention to incorporate Detrich's (1999) most salient contextual variables. (see Table 3.3). In addition, Caroll et al. (2007) developed a new conceptual framework that is measured empirically for understanding and measuring the process of intervention implementation. They conclude that this framework provides a means of measuring this variable. The authors have built on the five variables for fidelity of implementation proposed by Dusenbury et al. (2003). (see Table 3.4)

Table 3.3

Detrich (1999) Contextual Variables

Contextual Variable	This Study
Characteristics of the Child	• Pilot tested two questionnaires about Children's Test Anxiety to see what worked in the classroom. • Researcher changed the word anxiety to attitude and other American style words into U.K. English e.g., poorly, which can mean unwell in the U.K. to badly.
Resources required by the intervention	• Simple photocopying and administrative costs (Albin et al., 1996) offer of reimbursement. • Resources that are typically available (Wacker et al., 1990)
Similarity of Recommended Intervention with Current Practices.	• Practices in the classroom define a culture. (Albin et al., 1996) • The intervention was in the form of worksheets that the teacher could hand out as homework or discuss. • Homework is a usual routine. However, the worksheets were not marked as I could not increase teacher workload. The worksheets were meant to act as a discussion point.
Classroom cultures	• The intervention was designed to be easily deliverable and used discussion as its medium. • The Teacher's Manual recommended short, easily available resources for teachers to access if they needed further information such as Bob Bjork's video and I was ever present throughout by email, phone. • Two of the teachers had experience in delivering interventions before. Shapiro (1996)

Table 3.4

Caroll et al. (2007) Conceptual Framework

Elements of Implementation Fidelity	This Study
Adherence - Content	• Clearly defined. • All questionnaires and worksheets provided. • Simple instructions provided. • Teacher and student questionnaires provide feedback.
Coverage	• Teacher questionnaire gauges how much of the intervention was covered.
Frequency	• Once a week for fifteen to twenty minutes.
Duration	• Six weeks plus pre and post-test questionnaires and teacher and student feedback.
Moderators – Intervention Complexity	• Simple to deliver. • Some basic understanding needed. Resources provided for this.
Facilitation Strategies	• Is everyone receiving the same training and support? (Caroll et al., 2007). The Teachers' Manual was given to everyone. • I met with the teachers or headteachers in the schools. I communicated regularly with the teachers via email. • Teachers' responses encouraged; chance to make a difference. • Analysis of data provided where desired.
Quality of Delivery	• Researcher worked with all teachers and provided regular contact, opportunities for discussion. • Difficult to measure.
Participant Responsiveness	• This age-group, students tend to be keen. • Small gifts were offered if teachers thought it was appropriate. • These responses will depend upon teacher take up of the intervention too. In my experience, children of this age tend to value their teachers. • Student questionnaire provided but could be done individually or through a conversation with the teacher. Only one school did the questionnaires individually, the rest sent in notes that were summarized by the teacher.

As this experiment was a design-based theory experiment that was delivered in the field and conducted in primary classrooms and not conducted in laboratory conditions, it is important that the existence of extraneous variables is considered: participant factors, intervention factors and situational factors. (Cohen et al., 2017).

In this study, to try and manage the participant, intervention and situational factors, a manual (see Appendix B) was created and given to all the teachers who delivered the intervention. I included instructions for the administration and the rationale behind the study. I did emphasise one key point in the guidance.

> *'You must make the link between what you are doing in class (practising tests) and the cognition of the testing effect. My phrase is 'your testing routes are well oiled now' in preparation for the tests'*

As a teacher I know that schools are real world places and students are all individual. I have always loved my job as a teacher for the reason that no two days are the same. As such, I am aware that I could not control for every potential variable that may have occurred during the administration of the intervention.

I tried to ensure the conditions and delivery of the intervention were uniform. I was not able to deliver the intervention to all of the schools myself. Therefore, there could be a variation in how the intervention was delivered.

Post-positivism is the context of my educational research and I acknowledge the probabilistic and changing nature of the world (Cohen et al., 2017). As such, I acknowledge too, the bias that I may have as a researcher for the two groups, where I delivered the intervention. To measure this possible effect, I compared the schools where I had delivered the intervention with similar schools where the intervention was delivered by another teacher. (see Research Question 4). In addition, Appendix D details the feedback that I received from teachers and students from the different schools involved in the intervention. These results provide valuable feedback from professional colleagues and student voices and as such are of the utmost importance to me as both a teacher and researcher.

The schools that I approached were given a choice about whether to conduct the intervention or just to complete the questionnaires or to not participate. For some schools, completing the questionnaires was preferred and then, the intervention was offered later. Goldacre (2013) discusses the myths around randomized control trials in schools and concludes that they are ethically acceptable. This option is necessary because an intervention

may not benefit students. It might not work. Although, I am not using a randomized control trial, I have been able to offer an option to schools and the benefit of adopting the intervention for the following year is that the results will be known.

There were nine schools in total, two of which were used over different years, creating a total of 11 groups. For the schools that participated in consecutive years, they were in the intervention group the first year and the control group the second year, so with different participants and always Year 6.

3.2.1.1 Internal Validity. The design of the quasi, field type, experiment is a pre-test, post-test, control group design and this design controls for the effects of history, maturation, testing, instrumentation, and experimental mortality because there is a control group (Mertens, 2015). All the schools were in different regions and none of the schools were aware of each other, so I was able to avoid any inter group comparisons. Field and Hole (2003) suggest various threats to internal validity.

3.2.1.2 External and Ecological Validity. Tables 3.5 and 3.6 show how I have tried to minimize threats to external validity and Table 3.7 shows how I have tried to fulfil the criteria for Bracht and Glass's (1968) ecological validity.

Table 3.5

Field and Hole (2003) Threats to and Management of External Validity

Threat	This Study
Over-use of special participant groups	There is no extant research about using *the testing effect* to increase test taking self-efficacy to reduce test anxiety in 10-to 11-year-olds.
Restricted numbers of participants.	GPower*3.1 was used to establish power.

Table 3.6

Field and Hole (2003) Threats to and Management of Internal Validity

Threat	This Study
Group threats	• All participants are in Year 6 and aged between 10-to 11-years-old. • Intervention groups, all students attended independent or private school-similar financial backgrounds.
Regression to the Mean	• Participants with high overall scores were included in a separate analysis.
Time threats	• Timings designed to fit around the key testing phases in the schools.
History	• Pre-test questionnaires were taken weeks in advance of any high stakes' tests. • Post-test questionnaires were taken as soon as possible after the high stakes' test resulting in varied post-test timepoints.
Maturation	• The pre-test and post-test measures were taken within a 5-month framework.
Instrument change	• Adapted version of a reliable measure used (Children's Test Anxiety Scale, Wren & Benson, 2004) and exploratory and confirmatory factor analysis and measurement invariance on the adapted instrument conducted. • New instrument for measuring self-efficacy in test taking - exploratory and confirmatory factor analysis and measurement invariance.
Differential Mortality	• Given the age of the students, there was no 'drop out' rate • Students completed the pre and post-tests and intervention in school.
Reactivity and experimenter effects	• ANOVA to test for experimenter bias • The word 'anxiety' was changed to attitude in the questionnaire. • Different questionnaire tried with the same age-group in the pilot.

Table 3.7

Bracht and Glass Ecological Validity (1968)

Bracht and Glass Ecological Validity	This Study
Explicit Description of Experimental Treatment	• Teacher Manual (Appendix B)
Multiple-Treatment Interference	• Only one intervention was available.
The Hawthorne Effect – (just taking part in a study makes one feel special)	• Delivery of the pre-test questionnaires weeks in advance of the high stakes' tests but post-test could be an issue.
Novelty and Disruption Effects	• New treatment - may have produced positive results because it is novel - not have produced positive results because it has not been assimilated into a system yet.
Experimenter Effect	• An analysis was run between my groups and other groups in the intervention group to examine this effect-Research Question 4.
Pre-test Sensitization	• Pre-test questionnaires were taken weeks ahead of the high stakes' tests to avoid any sense of priming of test anxiety.
Post-test Sensitization	• This aspect is difficult to gauge. Post-tests were mainly completed within a week of finishing all the high stakes' tests, but the timing was different for different schools.
Interaction of History and Treatment Effects	• Contextual factors cannot be replicated from school to school – this aspect is the nature of classroom research.
Measurement of the Dependent Variable	• Instruments of measure and how they were used to measure are detailed in the thesis. • A tried and tested measure for test anxiety was used (Wren & Benson, 2014) but adapted. • New instrument created to measure Self-efficacy in Test Taking. Exploratory and Confirmatory factor analyses defined the sub-scales for measurement and their validity established. One limitation is that the new and adapted instruments have not been tested on a validation sample. Measurement invariance was conducted for all subscales used in all conditions.
Interaction of Time of Measurement and Treatment Effects	• The timing of the post-test may influence results. • Teachers asked to complete the post-test within a week of the last high stakes' test but this varied. In the schools where I was the teacher/researcher the post-test was taken directly after the high stakes' test on one occasion and within two days on the other occasion.

3.2.2. Effectiveness and Efficacy

This study was neither an effectiveness nor an efficacy trial (e.g., Flay, 1986; Marchand et al., 2011). Singal et al. (2014) suggest that efficacy trials can be defined by delivery of an intervention under ideal and controlled circumstances (providers identify a target population, recommend the intervention and that participants accept and adhere to the intervention). The Teacher's Manual (Appendix B) aims to mitigate any extreme deviations from the administration of the intervention but as had already been discussed, field experiments in education can include many extraneous variables (Cohen et al., 2017). Effectiveness trials refer to the performance of an intervention under real world conditions (Singal et al., 2014). Effectiveness trials account for factors that may moderate an intervention's effect. (Singal et al., 2014)

I like the way that Singal et al. (2014) refer to the differences between effectiveness and efficacy trials as being a continuum rather than a dichotomy. An efficacy study maximises the likelihood of observing an effect from the intervention if there is one but as noted above the idea of an efficacy trial is difficult to manifest in reality. In this study, I sought to control as many variables as I could. I tried to lead with a sense of efficacy but as a teacher, I was only too aware that efficacy can be compromised in a myriad of ways, on any day and in any classroom environment.

The extent to which the findings from a study can be applied to another is called generalizability (e.g., Field & Hole, 2003; Mertens, 2015). The best measure of generalizability is empirical testing. (Field & Hole, 2003). I pre-registered this study on the OSF (Open Science Framework, https://osf.io/j4dn2/) and uploaded the two main data sets to be open and transparent. (e.g., Anderson et al., 2019; Chambers, 2017; Gonzales & Cunningham, 2015) The cognitive element of worry from the test anxiety construct and self-efficacy in test taking were measured empirically.

3.2.3 The Delivery of My Intervention

The delivery of the intervention was timed to coincide with a time when students were completing practice tests in the classroom in preparation for high stakes' tests. The intervention was delivered over six weeks with approximately fifteen to twenty minutes of any lesson at any point of the week taken for a discussion about the testing effect. Students were given the worksheets as homework and could complete them if they wished. It was not mandatory, but I asked teachers to encourage students to look and read through the

worksheets. The idea was that if they did not understand any of the ideas or words, they could ask in the next lesson in the time given to the discussion.

Each week recall of the testing effect was done through discussing the ideas on the worksheets and students' ideas about these. There was no marking (grading) involved. Simultaneously, due to the nature of the high stakes' tests, the students were completing practice tests, in class, in preparation for the highly evaluative situations they were about to encounter. These tests were in Math, English, Spelling, Punctuation and Grammar for Standard Attainment Tests (schools) and Math, English, Verbal and / or Non-verbal Reasoning for other schools. I emphasized one key phrase to the teachers, who were delivering the intervention to say to the students and that was '*your testing routes are well-oiled'.* (see Table 3.8)

3.2.4 A Mini Summary of the Intervention

Table 3.8

Mini Summary of the Intervention

Week	Activity
Week 1	Cognition Worksheet 1. Introduction to the characters Roddy (Henry Roediger) and Jeffrey (Karpicke), and Bob (Bjork). Students invited to think what 'Test Enhanced Learning – Taking Memory Tests Improves Long Term Retention' might mean and also read some short paragraphs about repeated studying vs testing and were asked which was more effective and encouraged to summarise what they had read.
Week 2	Introduction to other *desirable difficulties* of spacing and interleaving and then back to *the testing effect* and asked to say which they thought was more effective "study, study, study, test" or "study, test, test, test"?
Week 3	Recall of previous x2 weeks learning -simply, what could be remembered?
Week 4	Letter to Bob Bjork in response to his letter expressing concern about how students learn.
Week 5	Carol Dweck 'Not Yet' video and summary of main messages.
Week 6	Letter of advice to an anxious friend who is worried about an exam. Students are reminded that the friend has also learned about testing.

Note. The cognitive scientists were introduced as if they were characters in a story so that the 10-to 11- year-olds would engage more with the material

The next chapter details the Research Questions. Chapter 5 is the pilot study. Chapter 6 describes the method for the main study. In chapter 7, I explain the preliminary analyses I conducted to establish the subscales for measurement for both questionnaires, the Children's Test Anxiety Scale (Wren & Benson, 2004) and the Self-efficacy in Test Taking questionnaire. In addition, I explore the assumptions needed for further analyses. The results of these are presented in Chapter 8 and then discussed and evaluated in Chapters 9 and 10. The implementation of the intervention is evaluated in Chapter 10.

4. Research Questions

I have used the extensive research about the robustness of the testing effect (e.g., Chan, McDermott & Roediger, 2006; Roediger & Karpicke, 2006a, 2006b), test anxiety theory (e.g., Zeidner, 1998) and self-efficacy theory (Bandura, 1997) to create an intervention that is easily administered by a classroom teacher in short, weekly sessions over a period of six weeks that uses retrieval practice to teach the testing effect to 10-to 11-year-olds to try and increase self-efficacy in test taking and reduce worry about taking tests. The intervention and its implementation are evaluated in Chapter 10.

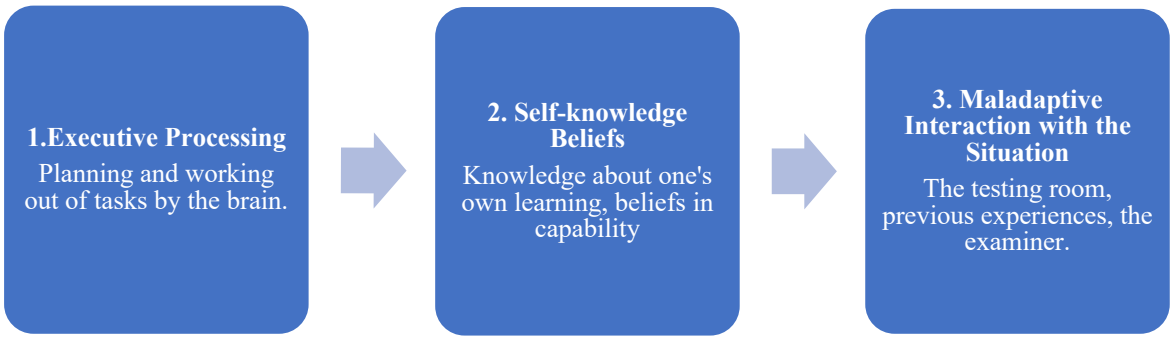
4.1 The Research Questions

There is a cognitive or worry element to test anxiety that was identified by Liebert & Morris (1967) and this is the element of test anxiety targeted by the intervention. The testing effect (e.g., Bjork, 1975; Karpicke & Roediger, 2008) has shown that "study, test, test, test" results in more effective retention of knowledge in the memory than "study, study, study, test" or "study, test, study, test" (Roediger & Karpicke, 2006 b). Students often misjudge what they know. These misconceptions can result in their mismanaging their learning and they would therefore benefit from knowing more about learning to learn. (Bjork et al., 2013). In this case, the students are taught about the testing effect at the same time as they were taking high stakes tests.

Test anxiety is a transactional construct dependent upon many proximal and distal factors (Zeidner, 1998) and will affect individuals differently depending upon *A state* or *A trait* (Spielberger & Vagg, 1995) types of test anxiety experienced. The S-REF (Self-referent executive function) model of test anxiety (Zeidner & Matthews, 2005) proposes three interactional areas of test anxiety (see Figure 4.1)

Figure 4.1

The S-REF Model - Self-referent Executive Function (Zeidner & Matthews, 2005)



It is the self-knowledge beliefs strand of the model that the intervention targeted in the hope that test anxiety was managed before it became a debilitating feature of the evaluative situation. The impact of the intervention was measured by the Research Questions.

Research Question 1: Does the intervention that uses retrieval practice to teach *the testing effect* reduce feelings of *worry* in 10-to11-year-olds?

I used an adapted version of the *thoughts'* subscale from the Children's Test Anxiety Scale (Wren & Benson, 2004) to measure the cognitive or worry aspect of test anxiety at pre and post-test timepoints. I adapted this measure to try and reflect changes in the test anxiety construct that suggest it is a multi-dimensional construct. (Putwain et al., 2020; von der Embse et al., 2021). There is a lack of research into test anxiety in 10-to 11-year-olds (e.g., Putwain, 2020) and research into using retrieval practice to manage test anxiety with this age-group is scarce (Agarwal et al., 2014; Barsham & Ellefson, 2020).

There are some notable exceptions to the general lack of research into test anxiety of this age-group and they are the meta-analyses about test anxiety interventions: Ergene (2003); Von der Embse et al. (2013); Weems et al. (2015) and Yeo et al. (2016). All of these authors have concluded that a variety of behavioural, cognitive and study skills interventions in test anxiety for elementary or primary aged students are effective. This study has focused on one element only, the cognitive element of worry.

As a classroom practitioner, I know that no two classrooms are the same and that the nature of a school-based intervention needs to be flexible, easy to administer and ideally,

require minimal input from a teacher. The intervention uses retrieval practice as the method of learning about the testing effect; the testing effect on the testing effect and its effects on increasing self-efficacy and reducing test anxiety are measured. To my knowledge this concept has not been researched before and with this age-group. The intervention and its implementation are evaluated in Chapter 10.

I expected to see a reduction in the mean value on the *Adapted Thoughts'* subscale of the Children's Test Anxiety Scale for the intervention group. The children learn about "study, test, test, test" of the testing effect in the intervention. They are told that their 'testing routes are well-oiled'. I expected that this cognitive knowledge would give a sense of mastery of self-efficacy in test taking to reduce uncertain control (Putwain & Pescod, 2018) and contribute towards perceived control (Putwain & Aveyard, 2018) to reduce worry. I thought that the result may be varied with the individual depending upon the *A state* or *A trait* nature (Spielberger & Vagg, 1995) of test anxiety.

Research Question 2a: Does the intervention increase feelings of self-efficacy in test taking in 10-to-11-year-olds?

The new instrument of Self-efficacy in Test Taking used to answer this research question is an original feature of this study and may add to the literature around the idea of increasing self-control in test taking by controlling the worry element of test anxiety (Putwain & Pescod, 2018) because it is a new measure that has been tried with 10-to 11-year-olds.

The aim is to equip students with mastery of belief in their test taking capabilities; belief in capability being a fundamental principle of Bandura's (1997) definition of self-efficacy and this case, is domain specific, to test taking, to combat anxiety in evaluative situations.

Self-efficacy theory (Bandura, 1997) shows that belief in capability improves performance, regardless of actual ability. Putwain and Pescod (2018) contribute to the evidence for school - based interventions being effective in reducing test anxiety through their experiment that explores whether a student's sense of being in control reduces test anxiety. Putwain and Aveyard (2018) link perceived control to the relationship between worry and performance and ask for interventions that focus on reducing worry and increasing perceived control.

As a teacher I have seen what happens to students who become anxious about taking tests. I witnessed a student, who was high in trait anxiety simply 'meltdown' in a mock, high

stakes' test. It was the S-REF (Self-referent Executive Function) model of test anxiety proposed by Zeidner and Matthews (2005), that was used as a target model for the intervention. From this model, it was the middle strand of self-knowledge beliefs that was targeted by the intervention to increase self-efficacy in test taking beliefs to provide a way of managing the worry element of test anxiety because students have positive self-beliefs to draw upon.

I expected to find that students, who completed the practice testing as a matter of classroom routine and who were then taught the cognition of the testing effect would feel as if they had mastery and understanding of their ability to take tests because they have been given the "why" (e.g., Zepeda et al., 2020) and that this understanding enabled them to have more belief in their capability and thus feel more self-efficacious in test taking situations; to have more control in the evaluative situation and thereby a reduction in anxiety. I think that student beliefs may be varied with the individual depending upon the *A state* or *A trait* nature (Spielberger & Vagg, 1995) of test anxiety. I expected that there would be a positive difference in feelings of self-efficacy in the post-test self-efficacy questionnaires that have been created for this study for the intervention group.

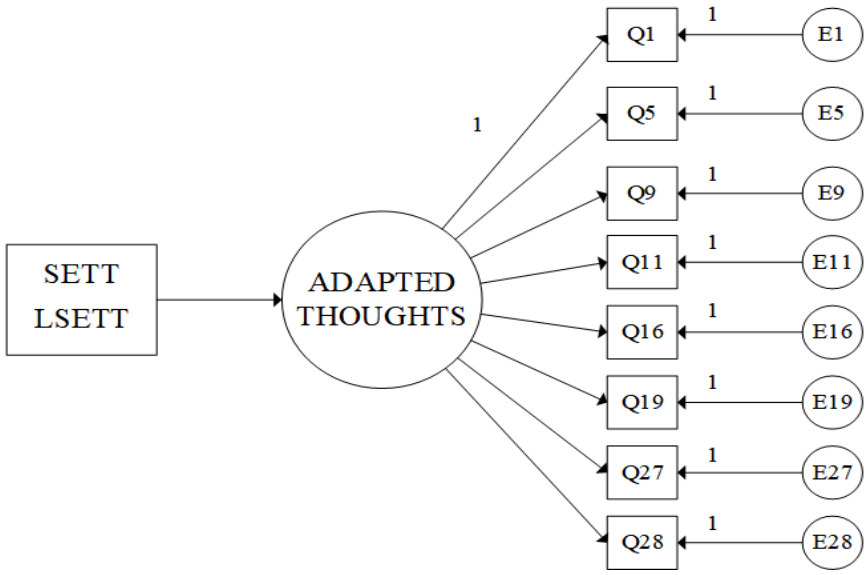
Research Question 2b: What do self-efficacy in test taking sum scores tell us about individual test anxiety?

I am not trying to attribute a causal effect to an individuals' self-efficacy in test taking to levels of test anxiety, but I thought it would be interesting to look closely at the relationship. A key aspect of the literature review is the important idea of self-efficacy and its role in the management of test anxiety because of the control it affords an individual (Putwain & Pescod, 2018; Putwain & Aveyard, 2018; Schunk & Pajares, 2005; Smith, 1989).

Through the intervention I tried to increase self-efficacy in test taking using self-efficacy theory (Bandura, 1997) and cognitive theory of the testing effect (e.g., Bjork, 1975; Roediger & Karpicke, 2006b) and I tried to explore this idea further by using MIMIC (Multiple Indicators and Multiple Causes) models to relate the covariate of self-efficacy in test taking to the worry element of test anxiety. Figure 4.2 demonstrates how the exogenous variable of a self-efficacy covariate is affecting the worry element of test anxiety measured by the Adapted Thoughts' subscale.

Figure 4.2

MIMIC (Multiple Indicators, Multiple Causes) Model of Self-efficacy Covariate in Relation to the Adapted Thoughts' Subscale



Research Question 3: Do feelings of *worry* predict how well student self-efficacy changes in the intervention group?

Putwain, Chamberlaine et al. (2014) suggest that their intervention (STEPS, Strategies to Tackle Exam Pressure and Stress) reduced test anxiety with 15-16-year-olds, who were high in test anxiety. The intervention was only given to the students who fulfilled the criteria for being high in test anxiety. Putwain et al. (2020) and von der Embse et al. (2021) have created a new multidimensional scale to measure test anxiety that may be used to determine recipients for a test anxiety intervention as no recent measure exists. However, the students in my study are younger and because of my experience as a teacher and knowledge from the literature review suggesting blanket test anxiety interventions can benefit the well-being of all students (Weems et al., 2015), the intervention was administered universally and not just to students who were high in test anxiety.

There is very little research into this age-group and test anxiety, and this is the first research question of its kind, to my knowledge, that examines the impact of a cognitive intervention that gives the 'why' testing works and uses the phrase 'your testing routes are

well-oiled' to 10-to 11-year-olds, who are high in test anxiety. The number of students who may experience high test anxiety is between 10-20% in any classroom. (Putwain & Daly, 2014). There are 31 students in this sample (18%) who are high in test anxiety as determined from their worry scores on the Adapted Thoughts' subscale. It is their data that is used to answer this research question.

I expected to find that these students who are high in test anxiety in the intervention group when compared to the control group, benefited from the intervention, and experienced an increase in self-efficacy in test taking feelings and a reduction in the cognitive element of worry of test anxiety. I expected that the worry scores for students who have high test anxiety in the control group remain at approximately the same level at post-test as pre-test.

Research Question 4: Are There Any Differences in the Reduction of Test Anxiety Within Researcher vs Teacher Groups from the Intervention Group?

As a researcher and teacher, I was very aware of experimenter bias (e.g., Bracht & Glass, 1968; Field & Hole, 2003) and fidelity of the intervention (e.g., Dusenbury et al., 2003). In Chapter 3, there are details of the steps I took to try and mitigate any experimenter bias in the results. Chapter 10 evaluates the success of these steps. This research question was included to examine if there was any difference between the groups where I delivered the intervention and where another teacher delivered the intervention. Although, I had taken steps in the design on the intervention to try and ensure that it was easily deliverable, in a primary classroom, by another teacher, I compared the two schools where I had delivered the intervention (one co-educational and one all girls) with similar groups where another teacher had delivered the intervention (one co-educational and one all girls) and all schools were in the independent sector of education so there were no free school meals to consider as a proxy for low income.

I expected that there would be little difference between the schools where I delivered the intervention and the schools where I didn't because I talked to the teachers or headteachers involved in the intervention schools and provided a manual for the delivery of the intervention (see Appendix D). However, I was keen to test this theory, hence the research question in case my knowledge as a researcher and the reality that I had delivered the pilot would affect the students differently.

Research Question 5: The SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) Question. Is There a Difference in Feelings of Test Anxiety Between Students Who Sit High Stakes' Paper Tests or Online Tests?

This last question was added after my data collection was changed due to the pandemic. I was waiting for one last set of control group data from a school that was taking SATs (Statutory Assessment Tests) when schools were closed due to the pandemic and the high stakes' tests were cancelled. I was contacted by a schoolteacher who had seen my article (Barsham & Ellefson, 2020) and who was interested in the research. Her Year 6 class were taking online high stakes mock tests and she was happy to complete the questionnaires with a view to delivering the intervention the following year.

I thought it would be interesting to see if there was any difference between taking high stakes tests online (at home) or in an exam room, as I am interested in the possible effects that an exam room may have with being able to recall exam content, for students who have test anxiety (e.g., Imundo et al., 2020; Bjork & Bjork, 2011).

I expected that there might be a difference between students who took real high stakes' paper tests and students who took mock, online tests with students who took paper tests in exam rooms experiencing more worry than students who took mock online tests at home.

All results in this thesis are exploratory as the research is making a novel contribution. The new subscales used to measure the impact of the intervention need to be cross validated on another sample.

5. Pilot Study

5.1 Purpose

Girls are more prone to the worry or the cognitive component of test anxiety (Hembree, 1988). Test anxiety can affect children, who are as young as seven. The purpose of the pilot was to introduce 10-to11- year-olds to the cognitive psychology of the desirable difficulty of the testing effect and to use the testing effect, through recall of the cognitive strategy itself, to increase self-efficacy prior to a high stakes' entrance examination.

5.2 Method

A quantitative approach was employed in the use of the Children's Test Anxiety Scale (Wren & Benson, 2004) with analysis of pre and post-test questionnaires through descriptive and inferential statistical analysis of the sample, $N = 11$. A small part of the experiment used student interviews that were carried out five days post-test.

Test anxiety was measured quantitatively. Thus, this experiment used mainly quantitative measures. In addition, there was a small, robust qualitative coding of interviews; these interviews sought to add to the measure of the individual, complex and transactional nature of the test anxiety.

5.2.1 Participants

There were eleven students in this study aged 10 -11 years - old, in their final year of primary education and with a mean age of 10.6 as of October 1st, 2017. The students were from an independent, all through girls' school. As a fee-paying school, the students were from an income bracket where parents are often both employed in professional capacities such as: doctors, dentists, and lawyers.

The students take assessments in Mathematics, English, and Non-Verbal Reasoning to try and gain a place at senior school (s). There are also academic scholarships available for Year 7 (11-to12-year-olds) that take the form of a reduction in fees.

5.2.2 Pre and Post-test Measures

The pre-test was delivered in the first week of November approximately nine weeks before the actual high stakes' test to avoid a sense of priming for anxiety.

Cassady and Johnson's 'Cognitive Test Attitude Scale' (2002) questionnaire was administered at the same time as the Children's Test Anxiety Scale, but the participants found the wording of the questionnaire difficult to understand. This questionnaire was not used in

the pilot other than for reference for one child who had missed out a response on the Children's Test Anxiety Scale on the thoughts' subscale and I was able to gauge her response by referring to a similar question from the Cassady & Johnson's (2002) questionnaire.

The study used the same pre-test and a post-test questionnaire about test anxiety to 11 participants. The questionnaire uses a Likert scale and is Wren and Benson's (2004) Children's Test Anxiety Scale. Descriptive and inferential statistics were obtained for the 30 different variables in the questionnaire and for the subscales of *thoughts, off task behaviours,* and *autonomic reactions* in both the pre and post-test questionnaires in the sample (N =11). (see Appendix F). The validity of the Children's Test Anxiety Scale questionnaire is established. Wren and Benson (2004) record Cronbach's *alpha* on the 30 items instrument as 0.92.

In addition, an informal interview was held with some of the participants a few days after the high stakes' test and an initial attempt at coding the students' responses undertaken using Nvivo©.

5.2.3 Materials

Test anxiety was measured using the Children's Test Anxiety Scale (Wren & Benson 2004). The word 'Anxiety' was replaced with 'Attitude' to avoid the issue of priming the sample with thoughts about anxiety. In addition, the wording on Question five 'I think I am going to get a bad grade' was replaced by 'I think I am going to get a bad result'. At this age, the students are rarely 'graded' in the English education system. In addition, the wording on Question 27 was changed from 'I think about how poorly I am doing' to 'I think about how badly I am doing'. Some students may think that 'poorly' means unwell.

This instrument provides scores on three test anxiety subscales:

1. Cognitive or *worry*, called *thoughts* in the Children's Test Anxiety Scale.
2. *Affective physiological* or *autonomic reactions*
3. *Behavioural* or *off task behaviours*.

Thirteen items measure *thoughts* or the cognitive element (e.g., 'While I am taking tests, I worry I will do something wrong'). Nine items measure the subscale *autonomic reactions* (e.g., 'While I am taking tests, I feel warm'). Eight items measure *off task behaviours* (e.g., 'While I am taking tests I stare'). Participants responded on a four-point scale (1= almost never, 2 = some of the time, 3 = most of the time, 4 = almost always).

5.2.4 Procedure for Data Collection

Permissions were sought for the use of both the pre-test and post-test questionnaires and for the interviews, from the sample's parents ($N = 11$) in writing and orally from the sample themselves and the school's overall Headmistress. I was Head of Prep in an all through school and thus needed to seek further endorsement from the overall Headmistress. All identities were anonymised.

There are three sets of quantitative data to consider: Year 6 ($N = 11$) pre-test and pre-intervention Children's Test Anxiety Scale questionnaire responses. Year 6 ($N = 11$) post-test and post intervention Children's Test Anxiety Scale questionnaire responses. In addition, there was an interview, which was conducted five days after the high stakes' test with the sample ($N = 11$).

5.2.5 The Pilot Intervention

The intervention (see Appendix B) that was about the cognitive effects of the desirable difficulty of the testing effect was delivered by me, the class teacher, as homework, over a period of six weeks, prior to high stakes' tests. The students were completing practice tests simultaneously in English and Math and Verbal Reasoning. The intervention was based on the following research papers: Bjork & Bjork (2019); Bjork, Dunlosky & Kornell, (2013); Karpicke, Blunt & Smith (2016); Kornell, Hays & Bjork (2009); Roediger & Karpicke (2008); Roediger & Karpicke (2006a, 2006b).

The information was presented to 10 -to 11- year-olds in child friendly language (see Appendix B) in the form of a homework worksheet. The research was presented in the form of a story using the first names of the 'characters': Robert or Bob (Bjork), Jeffrey (Karpicke) and Roddy (Henry Roediger). Female characters were introduced: Carol (Dweck) and Elizabeth (Bjork). A dialogue was created through the narrative so that it seemed as if the authors were talking to the children. A homework that was completed very well, was the 'Letter in response to a letter from Robert Bjork'. (see Appendix F)

The lessons were a week apart and at the beginning of each lesson, the students were invited to recall what they had discovered from the previous week's homework. Each lesson began with a cognitive conversation about desirable difficulties. Each week, I encouraged the students to recall the previous week's cognitive knowledge. The testing effect: study, test, test, test (Roediger & Karpicke, 2006b) was used to recall the cognitive knowledge about desirable difficulties and especially the testing effect over six weeks, on a weekly basis. The

students continued to take practice English and Math and Non-verbal reasoning papers during this time.

In addition, the participants also answered questions about the high stakes' tests in a lesson, five days after the test. These responses provided some qualitative data, through an interview that is valuable additional information in understanding whether or not the intervention was successful and the nature of test anxiety.

Teacher input was minimal. Checking understanding of the cognitive homework sheet or clarifying was all. The participants were all girls in a selective school who have historically achieved well in previous class papers in English, Mathematics, Verbal and Non-Verbal Reasoning papers.

5.3 Results

Inferential statistics were calculated using SPSS. Thirty paired *t*- tests to compare pre and post-test sample means were calculated for each variable. (see Table F.1) The number of *t* tests increases the likelihood of a Type 1 error. Although, the results suggest that for some of variables in the Children's Test Anxiety Scale questionnaire, there is a significant ($p < 0.5$) effect as a result of the intervention, I am very aware that, because the sample size was so small, the data must be interpreted within that parameter. It is likely that a Type 1 error is present because of experimentwise error rate (Field, 2013). The probability of a Type 1 error can be calculated thus:

Familywise error = $1- (0.95)^{30} = 0.78$

Effect sizes were calculated for this pilot using Cohen's *d*. 0.20 = small effect, 0.50 = medium effect, and 0.80 = large effect.

A little MCAR test was run to establish that the missing data were random and confirmed. For one student, I was able to seek a similar answer to this *thoughts'* question from the Cassady and Johnson (2001) questionnaire. For another student, the cognitive or *thoughts'* response was missed, and I used the other *thoughts'* responses, from the Children's Test Anxiety Scale to give the value of 4 to the answer.

The 'Go to the bathroom' question was missed out by a student. I looked at the other responses of this participant within the *autonomic* subscale and was confidently able to insert a 1 in for the answer.

The validity of the questionnaire's subscales was calculated using Cronbach's *alpha* using the large external candidate sample $N = 33$. The results are recorded in Table 5.1.

Table 5.1

Pilot - Validity of Questionnaire, N=33

Subscale	Validity
Thoughts' subscale	Cronbach's *alpha* = .85
Off task behaviour subscale	Cronbach's *alpha* = .69
Autonomic reactions subscale	Cronbach's *alpha* = .69

Note. $N = 33$

5.3.1 Interviews

The participants ($N = 11$) were interviewed informally five days after the tests in their usual English lesson with the researcher (me) as their classroom teacher. The conversation felt natural, and I made written notes.

The comments from these interviews that took place five days after the tests allowed me to ask direct questions about whether the participants considered the cognitive intervention to be successful or not. The responses from the participants gave valuable information about test anxiety and the effectiveness of the intervention.

The highly test anxious nature of one student was evident simply from the language that she used, referring to the assessment room as a 'torture chamber' and references to her mind 'going blank'. In another part of the conversation, she refers to being 'freaked out'. Her language is highly charged with emotive vocabulary.

Common themes in the participants' conversation included: self-talk about cognitive techniques, feeling confident because of mastery through the testing effect of both practice papers, self-efficacious talk, growth mindset and knowing about cognition, feeling that the brain was ready, testing routes were well-oiled. The participants felt that learning about the cognition aspect was true, factual so that it could be believed. 'Yeh, it made me feel confident because it's proved, you know it's true and not a maybe, it actually works.'

5.3.2 Discussion

The purpose of the pilot was to establish whether a six-week intervention using the cognition and practice of the desirable difficulty, the testing effect reduces the worry or cognitive element of test anxiety in 10 to 11-year-olds.

There were nine variables out of 30 that showed a significant result $p < .05$ after the intervention: I wonder if I will pass $p < .02$; My heart beats fast $p < .00$; I feel nervous $p < .05$; It's hard for me to remember the answers $p < .04$; My belly feels funny $p < .04$; I check the time $p < .03$; I think that I should have studied more $p < .02$; I look at other people $p < .03$; I feel scared $p < .04$.

Twenty-five out of 30 variables showed a positive effect of the intervention as the mean altered positively from pre to post-test in relation to the Likert scale. Sometimes, although the positive effect is not significant, it is consistently positive. The positive effect of the intervention was supported when the effect size was analysed and also from the interviews. However, the number of *t* tests increased the likelihood of a Type 1 error and thus the results must be considered preliminary.

These preliminary results suggest the intervention was a positive one in 25 out of 30 of the variables measured. The experiment and subsequent analysis of results is very limited because of the sample size $N = 11$.

The effect sizes were calculated using Cohen's *d* and if we apply the criteria, we can calculate that small effect, medium effect, and large effect. The effect sizes suggest that the intervention had a positive effect on all but one of the thoughts' subscale (what parents will say). The intervention was designed to reduce cognitive worry through increasing self-efficacy and was measured using the three subscales of the Children's Test Anxiety Scale.

I am encouraged, whilst acknowledging the limitations of the pilot study, by the significant results for the thoughts' subscale that suggests that the participants benefited from a reduction in: 'wondering if I will pass', 'finding it hard to remember' and the benefits of knowing they 'studied well', in this particular evaluative situation.

The interview supported the quantitative results through the positive comments about self-talk around "study, test, test, test", the benefits of practice papers and the belief in cognitive strategies being believable. This outcome is important because the actual research papers were used in the intervention but in Y6 friendly speak; Y6 were a critically friendly mass who respect authenticity. For results (see Appendix F)

5.3.3 Limitations

The sample size was very small ($N = 11$). The number of *t* tests increased the probability of a Type 1 error. The test anxiety questionnaire relied on self - report. The post-test questionnaire was delivered directly after the high stakes' test. There were some missing data; three responses were not completed.

First attempts at coding using Nvivo© indicated that the intervention had been effective in reducing the worry element of test anxiety. However, there is only one interview for the sample.

A further limitation was in not being able to access performance data for the high stakes' tests. Some general information was available such as the number of academic scholarships awarded (4 out of 13).

The question about 'checking the time' that is included in the off-task behaviours' subscale had the highest mean of 3.72 that reduced to 3.18 after the intervention. It is possible that this question is misleading because candidates are taught to check the time in examination situations, as part of good exam practice and to do so is not really an off-task behaviour. Self-efficacy in test taking was not measured in the pilot study and this measure will be useful to obtain in the main study.

5.3.4 Lessons Learned from the Pilot

The issue of sample size has been addressed and G Power© calculated as an a priori exercise before the main analyses were run. Furthermore, a control group was used for both the Children's Test Anxiety Scale measure and the Self-efficacy in Test Taking measure. Therefore, an increase in sample size, the use of a control group and the design of a further instrument have enabled me to use ANOVAs to measure the differences in group means in the main analyses.

During my upgrade viva, I realized that my study could be more effective if I measured the self-efficacy in test taking of the students, as well as levels of test anxiety. As no previous measure exists for self-efficacy in test taking for this age-groups, I designed a Self-efficacy in Test Taking instrument. Its validity was tested, and exploratory and confirmatory factor analyses used to determine the subscales. These subscales were then used as a measure of test taking self-efficacy in 10-to11-year-olds.

Given that the intervention's design is based on cognitive theory of retrieval practice, in this case of the desirable difficulty, the testing effect, I decided to use one subscale, not

three from the Children's Test Anxiety Measure (Wren & Benson, 2004) to measure the cognitive or worry element of test anxiety in 10-to 11-year-olds.

In addition, I had the opportunity to discuss the intervention with the Year 6 form teacher whose students were in the pilot and a Year 4 (8-to 9-year-olds) teacher. Both teachers felt that most of it was readily accessible for the 10-to 11-year-olds but suggested Week 2 the desirable difficulties worksheet was a little bit confusing and might need further clarification. As a result of this conversation, I added some clarification and signposts to Week 2 of the intervention. The intervention may need further review (see Chapter, 10, Reflections)

Furthermore, I decided to concentrate on quantitative analysis only as I had two data sets in the main study and because I had created two new measures, I knew the preliminary and main analyses would be extensive. I have included some feedback from students and teachers (see Appendix D) but have not analysed this qualitatively.

The pilot afforded me the opportunity to become familiar with the Children's Test Anxiety Scale measure (Wren & Benson, 2004). Knowing that the initial construction of the scale had seen a high correlation between the autonomic reactions' subscale and the thoughts' subscale in the validation sample, I took the unusual step of using both exploratory and confirmatory factor analysis to construct an Adapted Thoughts' subscale with which to measure the cognitive or worry element of the test anxiety construct in my data. In addition, changes to the test anxiety construct over time (e.g., Putwain et al., 2020; von der Embse et al., 2021) suggested exploring the factor structure would be a salient step.

5.3.4.1 Effect sizes. The effect sizes in the pilot were used to determine the sample size needed for the main study using G*Power 3.1. G *Power© was used to calculate the sample size for the mixed 2 x 2 ANOVAs. The power was set to 0.08 and *partial eta squared* values relative to sample size determined. For Research Questions 1, 2a and 2b, 3 and 4 for the ANOVAs, I input an 'A priori, *F* tests, ANOVA, Repeated measures, within -between interaction' using the medium effect size of *partial eta* = .06 and a power of 0.80, *alpha* = .05 and G*Power 3.1 calculated I needed $N = 24$.

For Research Question 2a and the ANCOVA, I input 'ANCOVA fixed effects, main effects and interactions.' I used a power of 0.80 and *alpha* of .05 with a medium effect size of *partial eta* = .06 and G*Power 3.1 calculated I needed $N = 125$.

For Research Question 5, I used G*Power 3.1 to calculate the sample size for an independent *t* test, inputting '*t*.tests, difference between two independent means (two groups)' and calculated the effect size from the means at $d = 1.40$ using a power of 0.80 and G*Power 3.1 calculated I needed $N = 30$ ($n = 15$ in each group). (see Table 5.2)

Table 5.2

Sample Size Calculations for the Main Study

Research Question and Type of Analyses	G*Power 3.1 Calculation	Sample Size in This Study
1 ANOVA	$N = 24$	$N = 170$
2a ANOVAs	$N = 24$	$N = 145$
2a ANCOVA	$N = 125$	$N = 145$
2b ANOVA	$N = 24$	$N = 145$
3 ANOVAs	$N = 24$	$N = 31$
4 ANOVA	$N = 24$	$N = 62$
5 Independent *t* test	$N = 30$	$N = 51$

These power analyses were used to verify the sample size for the main study. The sample sizes for all of the planned analyses are above the required sample sizes for this project. The next section, Chapter 6 provides my ontological and epistemological viewpoint and explores the design of my research. In addition, I provide a traditional APA method section and the ethical considerations for my study.

6. Method

This chapter is a hybrid of the format used by educational researchers and cognitive scientists. I start the chapter by considering the ontology, epistemology and methodology of my study that is placed in between education and cognitive science. I will then follow this consideration with the traditional APA method section.

I acknowledge the social reality (Blaikie, 2010, p. 92) that relies upon ontological and epistemological assumptions. The ontological position of cautious realism (Blaikie, 2010) and its accompanying epistemological partner of falsification (Popper, 1972a) is a good place to start to identify the branches of philosophy behind my study but with necessary qualifications.

In epistemological terms, my stance is postpositivist and thus I do not take a strict view of falsification (Wilson, 2017, p.299) but understand Popper's scientific intent. However, as a researcher of cognitive science and education, I acknowledge the fallibility of scientific knowledge. (Swann, 2003). Moreover, psychology lives exclusively under the post-positivist umbrella (Ellefson et al., 2019)

As I am measuring latent constructs in my study, it may be useful to consider the position of Guyon et al. (2018) who suggest that objects analysed in psychology differ fundamentally from physical objects and need an epistemology that is not empirical realism. They propose that measurement in psychology (psychometrics) is a type of pragmatic realism where "mental attributes are considered as real objects (linked to the brain) but set in a social praxis", Guyon et al. (2018, p.150). Mental attributes are defined as "characteristics (stemming from mental processes) that are inferred from the behaviour of a person" (Guyon et al., 2018).

In this study, I use cognitive psychology for the design of the intervention and pre and post-tests to measure the latent constructs of worry and self-efficacy in test taking. Guyon et al. (2018) suggest that pragmatic realism is where the equilibrium of an individual is considered as a psychological attribute and is a reality that needs to be analysed in a dynamic relationship with the social environment. This idea potentially links to the theories used in this study: the social cognitive theory of self-efficacy and dynamic of reciprocal causation (Bandura, 1986b) and the interactional dynamics that are found in the S-REF model of test anxiety (Zeidner & Matthews, 2005). Ellefson et al. (2019) suggest that the post-positivism is assumed and not questioned in psychology and that use of the experimental method aims to

reduce the bias of the researcher, but they also acknowledge that criticism of post-positivism can be common in education.

6.1 Method – Rationale

This design experiment (Cohen et al., 2017) has its roots in quasi-experimental research. This experiment has avoided the artificial world of the laboratory and was designed to have practical relevance to the world of teaching (see Chapter 9 and Chapter 10 for an evaluation).

The design experiment is rooted in the postpositivist paradigm (e.g., Cohen et al., 2018; Mertens, 2015; Panhwar et al., 2017). The purpose of the research from a postpositivist perspective is to embrace the scientific method but to recognise that there is no absolute truth (Cohen et al., 2017). In the postpositivist paradigm there is a pluralism between positivist and interpretivist epistemologies (Phillips & Burbules, 2000). Quantitative analyses are used in this study (see Chapter 8). A small amount of teacher and student feedback was collected (see Appendix D) The results of the design experiment should show whether or not a cognitive intervention using the testing effect can increase self-efficacy and reduce the worry element of test anxiety and if the results can be generalized to a school population.

The method is about the data collection and is not an evaluation of the intervention. Chapter 3 details the intervention and how it was administered, and Chapters 8, 9 and 10 evaluate the impact of the intervention.

6.1.1 Participants

Data were collected from 170 participants (males $n = 62$, females $n = 108$), from nine primary schools for the Children's Test Anxiety Scale questionnaire. All students were in Year 6, the final year of primary education. The mean age of participants in both of these samples was 10.5 years ($SD = .50$).

Two of the schools were used twice over two different years, making a total of 11 sets of school data. Eight of the nine schools were independent schools and fees are paid for the students to attend. One school in the control group was a state school and there were ten students who received free school meals. Four schools took high stakes' SATs (Standard Assessment Tests), five schools took other high stakes' independent school exams, one school took an online high stakes' SATs (Standard Assessment Tests) type of exam and another school took a high stakes' mock SATs (Standard Assessment Tests). Mock high stakes tests were due to the cancellation of SATs (Standard Assessment Tests) as a result of

the pandemic. There were 85 participants in the group that received the intervention (males *n* = 28, females *n* = 57) and 85 participants in the group that did not receive the intervention (males *n* = 34, females *n* = 51).

For the Self-efficacy in Test Taking instrument, data were collected from seven schools. There were 145 participants (males *n* = 48, females *n* = 97). There were 59 participants in the intervention group (males *n* = 14, females *n* = 45) and 86 participants in the control group (male *n* = 34, female *n* = 52). Six out of seven schools were independent, and fees are paid for students to attend. One school in the control group was a state school and ten students received free school meals.

6.1.2 Missing Data

Very little data were missing. For the Children's Test Anxiety Scale questionnaires, if a student missed out a response, I took the mode number for the same subscale. Where students had used two numbers, E.g., 2 /3, I used the mode for the same subscale. (see Table 6.1) There was one pre and post questionnaire that was unusable (Children's Test Anxiety Scale) in the control group as it had been completed with zeroes and zeroes did not exist as an option. One school had three self-efficacy questionnaires missing. Where I have detailed an item missing, it was only one item out of thirty.

G*Power estimates for sample size needed to run the different ANOVAs, ANCOVA and *t* tests to answer the research questions, were exceeded for all the research questions based on G*Power A priori calculations (details in section 5.3.4.1)

6.1.3 Materials

Two questionnaires were used at pre-test and post-test. The Children's Test Anxiety Scale (Wren & Benson, 2004) and the Self-efficacy in Test Taking instrument that was created by me. Exploratory and confirmatory factor analyses were used to establish a new cognitive or thoughts' subscale, the Adapted Thoughts' subscale for the Children's Test Anxiety Scale data that was used for the final analyses and exploratory and confirmatory factor analyses were conducted for the Self-efficacy in Test Taking instrument that resulted in two subscales with which to analyse the Self-efficacy in Test Taking data. A teachers' manual detailing the delivery of and resources for the intervention was provided. (see Appendix B).

Table 6.1

Schools, High Stakes Tests, Sample Number Used and Missing Data

School	High Stakes Tests	Intervention / Control	*N*	Missing Data
Co-ed	SATs	Control	*N* = 16	
Co-ed	SATs	Intervention	*N* = 15	CTAS x1 mixed response value.
Co-ed	Online high stakes'	Control	*N* = 30	
Co-ed	SATs	Intervention	*N* = 4	No Self-Efficacy data*.
Girls	Senior School Entrance Papers	Intervention	*N* = 20	CTAS x1 mixed response value.
Girls	Senior School Entrance Papers	Control	*N* = 15	CTAS x1 mixed response. CTAS x1 pre and post questionnaire unusable.
Co-ed	Senior School Entrance Papers	Intervention	*N* = 19	No Self-efficacy data*; CTAS x4 mixed response values; x2 CTAS items.
Girls	Senior School Entrance Papers	Intervention	*N* = 15	Three Self-efficacy questionnaires.
Co-ed	Mock SATs	Control	*N* = 4	
Co-ed	SATs	Intervention	*N* = 12	
Co-ed	Senior School Entrance Papers	Control	*N* = 21	

Note. Co-ed = Co-educational, * Self-efficacy questionnaire was being trialled at the time with my own classes. CTAS = Children's Test Anxiety Scale Data.

6.1.3.1 Test Anxiety. The cognition element or worry of test anxiety was measured using the Adapted Thoughts' subscale from the 30 item Children's Test Anxiety Scale instrument (Wren & Benson, 2004). This questionnaire (see Appendix A) comprises three subscales. Firstly, the thoughts or cognitive subscale (13 items) that measure the test anxiety element of worry. An example is 'While I am taking tests, I think about what will happen if I fail'. Secondly, the autonomic reactions' subscale (nine items), that measure physiological changes. An example is 'While I am taking tests my belly feels funny' and lastly the off-task behaviours' subscale (eight items) that measure attention deficits. For example, 'When I am taking tests, I tap my feet'. Participants responded to items using a 4-point Likert type scale (1 = "almost never", 2 = "some of the time", 3 = "most of the time", 4 = "almost always"). The Adapted Thoughts' subscale consisted of eight items. Seven of these items were on the original thoughts' subscale and Question 28 'When I am taking tests, I feel scared' was moved from the autonomic reactions' subscale to the Adapted Thoughts' subscale as a result of factor analysis.

6.1.3.2 Self-Efficacy in Test Taking. Self-efficacy in Test Taking was measured using a new instrument. (see Appendix A) I designed this instrument using guidance from Bandura's 'Guide for Constructing Self-efficacy Scales' (Bandura, 2006). This new measure, and its reliability and validity are assessed in Preliminary Data Analysis II.

Originally this scale was 15 items and, following exploratory and confirmatory factor analyses, the instrument resulted in two subscales: Self-efficacy in Test Taking subscale consists of nine items that represent positive feelings of efficacy about test taking. An example is 'If I find a test hard, I keep trying'. Participants responded to items on a 4-point Likert type scale, that was very similar to the test anxiety measure. (1 = "I never do this', 2 = "I do this sometimes', 3 = "I do this many times", 4 = "I always do this". High scores on the Self-efficacy in Test Taking scale represent a student with high self-efficacy in test taking feelings. The second subscale, the Little Self-efficacy in Test Taking has four items that represent negative feelings of efficacy about taking tests. An example is 'In tests, I panic if I see people finishing before me'. High scores on this scale represent a student with low self-efficacy in test taking feelings.

6.1.4 Procedures

Self-report data from both questionnaires were collected approximately 6 -8 weeks before the high stakes' tests for both the intervention and control groups. For some schools this month was October / November and for other schools this month was March /April,

depending upon the type of high stakes' tests involved. The same self-report data from both the questionnaires was collected as soon as possible after the high stakes' tests and the timings of this collection varied considerably depending upon the number and range of high stakes' tests being taken.

A Training Manual (Appendix B) was provided for all teachers. I visited or met via Zoom© with all the headteachers /teachers in schools, who agreed to deliver the intervention to talk through the training manual and to explain the letter for parental permissions. Teachers emailed me to confirm when they had received permissions and then administered the questionnaires at a time that was convenient to them. The same questionnaires were given to the students after all the high stakes' tests were completed and the timings of these varied. All teachers, who delivered the intervention, had my personal email and phone number in case of any queries. I sent regular emails checking on their well-being and progress and offering support and help throughout the duration of the interventions. Small rewards for participation were available for the students if teachers thought appropriate. Teachers were invited to complete a questionnaire about the intervention afterwards and also to ask the students' their opinions.

6.2 Design

Cohen et al. (2018, p. 402) refer to the design of pre-test and post-test with an experimental and a control group as a "true experimental" design. However, in this design, the groups were not selected by random sampling from the population and thus do not fulfil the criteria. In a quasi-experimental design, the experimenter does not have complete control over the manipulation of the independent variable (Field & Hole, 2013) This lack of control was the case in my study as I asked other teachers to deliver the intervention and different schools were involved. I did not have complete control over the timing of the measurement of the dependent variable either and so this aspect of a quasi-experiment has not been fulfilled.

Students were not randomly assigned to the intervention or control group as this assignment was practically impossible in this educational context.

My experiment is a design experiment with its roots in the quasi-experimental and this hybrid is considered to be a special case of a field experiment (Cohen et al., 2017). The roots are in experimental research with a pre and post-test measuring the impact of the intervention and the outcomes are for direct practical relevance and use in primary classrooms. The role of theory is important in the design of the intervention (Bradley and

Reinking, 2011) and the way that the intervention fulfils the criteria for a design experiment is explained more fully in Chapter 3.

Sum scores on the Adapted Thoughts' subscale from the Children's Test Anxiety Scale instrument that measured the cognitive or worry element of test anxiety were the dependent variable that was measured in Research Question 1. This question was answered using a 2 x 2 mixed factorial ANOVA. The intervention was the independent variable 'Group' (intervention and control groups) and 'Timepoint' was the repeated measure for pre-test and post-test. Multilevel linear analysis followed when the assumption for homogeneity of slopes prior to an ANCOVA was not met.

Sum scores on the Self-efficacy in Test Taking subscales were the dependent variables in Research Question 2a. The research question was answered using 2 x 2 mixed factorial ANOVAs for both subscales. Criteria for ANCOVA and multilevel linear analysis were not met for the Self-efficacy in Test Taking subscale. For the Little Self- efficacy in Test Taking subscale, the ANOVA produced an interaction and *Tukey's Honestly Significant Difference* post-hoc test was used to identify the groups involved. The independent variable 'Group' (intervention and control) was the between measure and 'Timepoint' was the repeated measure for pre-test and post-test. An ANCOVA was used for the Little Self-efficacy in Test Taking subscale as the assumption of homogeneity of slopes was met. The post-test score was the dependent variable, and the intervention or control group was the independent variable, and the centred pre-test score was the covariate.

Research Question 2b used Multiple Indicator and Multiple Causes (MIMIC) models with the Adapted Thoughts' subscale was the latent variable and the sum scores on the Self-efficacy in Test Taking Subscales was the regressed continuous covariate to see if there is a relationship between the two.

Research Question 3 measured students who were high in test anxiety using scores from the Self-efficacy in Test Taking subscales. Participants used for this analysis were determined by their scores from the mean of the Adapted Thoughts' subscale at the post-test timepoint. The dependent variable was the score on the Self-efficacy in Test Taking subscales and the independent variable was the group (intervention or control). Results were calculated by running a mixed 2 x 2 factorial ANOVA and *Tukey's Honestly Significant Difference* post-hoc test was used for the interaction. In addition, a mixed 2 x 2 factorial ANOVA was run to measure students who are high in test anxiety scores on the Adapted

Thoughts' subscale (dependent variable) with intervention and control groups as the independent variable.

Research Question 4 used the scores from the Adapted Thoughts' subscale as the dependent variable and the independent variable was the researchers' groups measured against another teacher with an intervention group. Results were calculated by running a 2 x 2 factorial ANOVA with timepoint as the repeated measure and the researcher led vs. non researcher led as the independent variable.

Research Question 5 used a comparison of between group means in the form of an independent *t* test. The dependent variable was the scores on the Adapted Thoughts' subscale from the Children's Test Anxiety Scale data and the independent variable was between the group who took high stakes' tests online and the group who sat paper tests.

6.3 Ethical Considerations

I met with headteachers or teachers for the schools involved to explain the research and obtained their agreement either to deliver the intervention or complete the questionnaires with the option of delivering the intervention at a later date, if they wished. For the schools where I was a headteacher, I sought the permissions of my line managers in the first instance to either conduct the research or to approach fellow headteachers in my network. Questionnaires and the intervention were administered by Year 6 teachers or known members of staff to the students.

I was headteacher at the schools where I delivered the intervention and happily experienced excellent relationships with my students. I am known to the students for being a caring and approachable headteacher who helps them in their school days. My door is always open, and students can (and do) wander in freely. I can reflect that students may want to please me as students of this age generally do want to please their teachers but also think that the rapport I was privileged to have with students would mean that if they were unhappy or felt awkward completing any aspect of the research, they would have told me when I asked them. I gained their consent as well as parental consent and students were enthusiastic and enjoyed the new learning.

When I met with teachers or headteachers, I discussed the requirement for parental consent and the importance of asking students themselves verbally if they were comfortable with the processes. I gave teachers / headteachers a copy of the Teachers' Manual and showed them the sample letter they should use (see Appendix E). Written permissions from

parents were obtained using active consent (opt in) via school email systems. Student initials were used on the questionnaires to make matching pre and post-tests easy. All data has been anonymized in the analyses and no student initials used. The project was registered with the Open Science Framework. (https://osf.io/j4dn2/) A completed ethics form and risk assessment were submitted to the University of Cambridge. The project was reviewed by the University of Cambridge ethics committee.

I have adhered to the BERA Ethical Guidelines for Educational Research and the APA Ethical Principles of Psychologists and Code of Conduct. I have always prioritized students' and colleagues' well-being. I have acted professionally and with respect to my colleagues involved in this study, mindful of their workloads and time commitments and have kept the students' potential beneficial outcomes at the heart of the design of the study.

7. Data Analytic Plan and Preliminary Analyses

This chapter provides an overview of the preliminary analyses and the assumptions that are considered prior to Chapter 8, the Results chapter. This chapter is a statistical preamble to the main analyses and consists mainly of exploratory and confirmatory factor analyses to ensure that I have obtained optimal measures for the final analyses.

7.1 Overview of Chapter 7

Part 1 of the preliminary analyses are factor analyses for the Children's Test Anxiety Scale data. Part II preliminary analyses, refer to factor analyses for the Self-efficacy in Test Taking data. In these sections, I detail my rationales for using exploratory factor analysis on a known measure. In addition, I discuss why I have chosen to use parametric tests with data that is from Likert type measures. Furthermore, all the decisions I made in the exploratory and confirmatory factor analyses processes are discussed along with information about the fit indices used. Preliminary analyses III refer to all other assumptions that need to be met prior to final analyses. Tables 7.1 to 7.7 detail the descriptive statistics, assumptions and analyses used for the five research questions.

Table 7.1

Descriptive Statistics for Children's Test Anxiety Data and Self-efficacy in Test Taking Data

Children's Test Anxiety Scale Descriptive Statistics RQ1, 4, 5	• Mean, Standard Deviation, Skew, Kurtosis and Standard Error for every item on the Children's Test Anxiety Scale instrument at pre-test and post- test (Timepoint) and for both intervention and control groups. (Group) • Mean, Standard Deviation, Skew and Kurtosis and Standard Error for the original thoughts' subscale in the Children's Test Anxiety Scale questionnaire at pre-test and post-test with intervention and control groups. • Histograms for Response Frequencies for the original thoughts' subscale at timepoint pre-test and post-test and for the intervention and control groups. • Violin Plots showing Response Frequencies for the Adapted Thoughts' subscale at pre-test and post-test with intervention and control groups.
Self-efficacy in Test Taking Scale Descriptive Statistics RQ2a, 2b, 3	• Mean, Standard Deviation, Skew, Kurtosis and Standard Error for every item on the Self-efficacy in Test Taking instrument at pre-test and post-test (Timepoint) and for both intervention and control groups. (Group) • Mean. Standard Deviation, Skew, Kurtosis and Standard Error for the subscales in the Self-efficacy in Test Taking instrument at pre-test and post-test with intervention and control groups. • Mini histograms for responses for individual items pre and post-test (x15 items) • Violin Plots of Response Frequencies for the pre-test and post-test timepoints and the intervention and control groups for the two subscales: Self-efficacy in Test Taking and Little Self-efficacy Scale in Test Taking.

Note N = 340, n = 85.

Table 7.2

Analyses and Assumptions for Research Question 1: Does the Intervention that uses Retrieval Practice to Teach the Testing Effect Reduce Feelings of Worry in 10-to11-year-olds?

Analyses and Assumptions	Details
Confirmatory Factor Analysis	• Tests whether the distribution of scores is significantly different from a normal distribution. (Field, 2012)
• Shapiro-Wilk Test: (Shapiro & Wilk, 1965)	• A significant value indicated a deviation from normality, but the test is affected by large samples. (Field, 2012)
• Sample size	• E.g., 10 subjects / variable (Nunnally 1978), 2 subjects / variable (Kline, 1994) but cannot be one size fits all (Wolf et al., 2013)– *see rationale in text.* (This study, N = 170, n = 85 intervention, n = 85 control, 13 items for the original subscale and 8 items for the adapted subscale).
• A priori model	• Children's Test Anxiety Scale (Wren & Benson, 2004)
• Random sampling	• Not met – difficult to achieve with classroom-based research (Cohen et al., 2013)
Exploratory Factor Analysis	• Represents the ratio of squared correlation between variables to the partial correlation between variables. (Field, 2012)
• Kaiser-Meyer-Olkin Measure of Sampling Adequacy (Kaiser, 1970)	• Value of between 0 and 1; a value close to 1indicates that the patterns of correlation are compact, and that the data would give clear and reliable factors from factor analysis. (Field, 2012)
• Bartlett's Test for Sphericity (Bartlett, 1951)	• Examines whether the variance / covariance matrix is proportional to an identity matrix. (Field, 2012)
• Correlation, multicollinearity, and singularity.	• There should be some correlation between the items, but the relationship should not exceed r = > .90 as this result would suggest multicollinearity. (Field, 2012)
2 x 2 Mixed, Factorial ANOVA	• 2 x 2 Mixed factorial ANOVA (with timepoint and group as the independent variables and the sum scores on the Adapted Thoughts' subscale as the dependent variable to measure the effect of the intervention)
• Levene's Test for Homogeneity of Variance (Levene, 1960)	• Tests the hypothesis that the variances in different groups (intervention and control) are equal. (Field, 2012)
	• A significant result indicates that the variances are significantly different, and this result would violate the assumption. (Field, 2012)
• Normality of Distribution	• Field (2012) – central limit theorem – the sampling distribution will tend to be normal regardless of the population distribution in samples > 30.

Table 7.2 (continued)

Table 7.2 continued

Analyses and Assumptions	Details
Latent Means Analysis	
• Measurement Invariance Testing Across Groups, Timepoints and Gender.	• (Brown, 2015; Finch & French, 2015)
Multilevel Linear Model	
• Homogeneity of variance, additivity, linearity, normality	• (Field, 2012)
• Homogeneity of slopes for ANCOVA is not met	• (Field, 2012)
• Coefficients have normal distribution of random intercepts in different contexts	• Default of variance components covariance structure is fine for random intercepts model (Field, 2012)
• ML estimator to compare models	• Field, 2012
• Multicollinearity	• Grand Mean Centring of predictors (Kreft & deLeeuw, 1998)

Note. Bold font indicates main analyses.

Note. N = 290, n = 59 intervention*; n = 86* control

Table 7.3

Analyses and Assumptions for Research Question 2a: Does the Intervention Increase Feelings of Self-efficacy in Test Taking in 10-to-11-year-olds?

Analyses and Assumptions	Details
• **Exploratory Factor Analysis of the Self-Efficacy in Test Taking instrument.** • **Confirmatory Factor Analysis of the Self-Efficacy in Test Taking instrument.**	• See Table 7.2 for details and assumptions • See Table 7.2 for details and assumptions
• **2 x 2 Mixed Factorial ANOVA**.	• 2 x 2 Mixed Factorial ANOVAs (with timepoint and group as the independent variables and the sum scores on the Self-efficacy in Test Taking subscale and the Little Self-efficacy subscale in Test Taking subscale, as the dependent variable, to measure the effect of the intervention) • See Table 7.2 for details of assumptions
• **Post Hoc Tests** (Little Self-efficacy in Test Taking Subscale) • Homogeneity of variance (Field & Hole, 2003)	• *Tukey's HSD* (*Honest Significance Test - more here*) if homogeneity of variance.
• **Latent Means Analysis** • **Measurement Invariance Testing Across Groups, Timepoints and Gender**	
ANCOVA **Assumptions** • As per ANOVA plus homogeneity of slopes (Field, 2012)	• See Table 7.2 for details of assumptions

Note. Bold font indicates main analyses.

Note. N = 145, n =59 intervention*; n = 86* control.

Table 7.4

Analyses and Assumptions for Research Question 2b: What do Self-efficacy in Test Taking Sum Scores Tell Us About Individual Test Anxiety?

Analyses and Assumptions	Details
• **MIMIC models (Multiple Indicators, Multiple Causes)** • CFA on Children's Test Anxiety Data (Goodness of fit statistics for latent factor) • Observed continuous covariate of the Self-efficacy in Test Taking Subscales.	• (Finch and French, 2015; Brown, 2015) • For the Adapted Thoughts' latent construct • Self-efficacy in Test Taking sum scores for both subscales

Note. Bold font indicates main analyses.

Note. N = 145, n = 59 intervention*, n = 86* control.

Table 7.5

Analyses and Assumptions for Research Question 3: Do Feelings of Worry Predict How Well Students Self-efficacy Changes in the Intervention Group?

Analyses and Assumptions	Details
• **2 x2 Mixed Factorial ANOVA**.	• 2 x 2 Mixed factorial ANOVAs (with timepoint and group as the independent variables and the sum scores on both Self-efficacy in Test Taking subscales and the Adapted Thoughts' subscale, as the dependent variable, to measure the effect of the intervention for students who have high test anxiety scores. • See table 7.2 for details of assumptions
• **Post Hoc Tests** (Little Self-efficacy in Test Taking Subscale) • Homogeneity of variance (Field & Hole, 2003)	• *Tukey's HSD* (*Honest Significance Test*) if homogeneity of variance.

Note. Bold font indicates main analyses.

Note. N = 31

Table 7.6

Analyses and Assumptions for Research Question 4: Is There a Significant Difference Between the Groups in the Intervention Group Where the Researcher Delivered the Intervention?

Analyses and Assumptions	Details
• **2 x2 Mixed Factorial ANOVA**.	• 2 x 2 Mixed ANOVAs (with Timepoint and Group as the independent variables and the sum scores on the Adapted Thoughts' subscale, as the dependent variable, to measure differences in the effect of the intervention for students who received the intervention from me and from other teachers. • See table 7.2 for details of assumptions

Note. Bold font indicates main analyses.

Note. *N = 62*

Table 7.7

Analyses and Assumptions for Research Question 5: The SARS-CoV-2 Question. Is There a Difference in Feelings of Test Anxiety Between Students Who Sit High Stakes' Paper Tests or Online Tests? (*N = 51, n = 30 online tests, n = 21 paper tests*)

Analyses and Assumptions	Details
Independent *t* test • Scores in different treatment conditions are independent because they come from different people (Field, 2012)	
• Homogeneity of Variance	• In theory we assume equal variances but in reality, we don't. (Field, 2012) Violating this assumption only matters if you have unequal group sizes – if you don't have unequal group sizes, the assumption is more or less irrelevant and can be ignored. *N* = 51 at pre-test and post-test (*n* = 30, online, *n* = 21, paper tests)

Note. Bold font indicates main analyses.

7.1.1 Evaluating the Self-report Questionnaires

Both questionnaires in this study use a four-point Likert type scale. Factor analyses are used to reduce the data and obtain a subscale with which to measure the cognitive element of test anxiety, worry, in the Children's Test Anxiety Scale data (Wren & Benson, 2004) and subscales with which to measure the Self-efficacy in Test Taking data.

For the Children's Test Anxiety Scale, an initial confirmatory factor analysis, followed by exploratory factor analysis, and then a final confirmatory factor analysis on all the data and by pre-test / post-test intervention and control groups, were carried out, to establish the Adapted Thoughts' subscale to measure the worry element of test anxiety. (see Tables G.13 -G.22 for exploratory factor analyses). Cronbach's *alpha* (Cronbach, 1951) is reported for the original subscale in all of the pre-test and post-test conditions and is acceptable (see Table 7.10). However, the decision to run further factor analysis was partly due to the results of the goodness and badness of fit statistics in the initial confirmatory factor analysis, especially the unacceptable *RMSEA* (Root Mean Square Error of Approximation) values. (see Table 7.9) *RMSEA* is a widely used and recommended index from the parsimony correction indices (Brown, 2015). *RMSEA* assesses how well a model fits reasonably well in the population. In addition, there have been developments in the test anxiety construct since 2004 as identified by Putwain et al. (2020) and their new Multidimensional Test Anxiety Scale.

I have used the *RMSEA* values as the main guide in assessing the fit of the proposed subscales. Values of 0 indicate a perfect fit and values near to 0, a good fit. Hu and Bentler (1999) recommend a *RMSEA* of 0.6 or less and *CFI* (comparative fit index) of 0.95 or greater where maximum likelihood is used as a good fit between the target model and the data. Browne and Cudeck (1993) suggest that *RMSEA* less than 0.8 is an adequate model fit and > 0.10 should be rejected. The *RMSEA* values for the Children's Test Anxiety Scale data are all > 0.10 for the initial confirmatory factor analysis on this data and thus the original subscale of thoughts is not a good fit to the data. (see Table 7.9)

Measurement invariance is then conducted after the final confirmatory factor analysis to examine the proposed subscale across pre-test and post-test (timepoint) groups, intervention and control groups and gender groups. If measurement invariance 'holds' then the proposed subscale is a good measure for the different groups that need measuring (e.g., Finch & French, 2015).

The main measure that I took was the difference in *CFI* (Comparative fit index) between the different sequential stages of the process. *CFI* is a well-behaved fit index (Brown, 2015). The difference was established using a comparison of goodness of fit indices between the models and moving to the next stage if the *CFI* values did not exceed more than 0.1(Chen, 2007) and verifying this result with probability of *chi-square* (*Pr > Chi Sqr*) statistic from an ANOVA that compared the two models. If the ANOVA results in a non-significant value for the probability of *chi-square,* then the two models can be deemed the same and measurement invariance is said to hold and the next sequential step does not need to be taken. If the ANOVA is a significant value, then the measurement invariance proceeds to the next stage: configural to constrained to metric to scalar.

For the Self-efficacy in Test Taking instrument, it is not possible to conduct an initial confirmatory factor analysis because it is a new measure and therefore there is no a priori study to draw upon. As with all new measures, it is appropriate to explore the number of factors through exploratory factor analysis (e.g., Putwain et al., 2020) although due to the sample size, the subscales have not been cross validated. This situation is a limitation of the process. Confirmatory factor analysis followed the exploratory factor analysis (e.g., Fabrigar et al., 1999).

Factor analysis is the most commonly used evaluation for multiple item testing instruments such as questionnaires. It is also used for construct validation as demonstrated in the original search for a new test anxiety measure (Wren & Benson, 2004) and more recently in the development and validation of a new Multidimensional Test Anxiety Scale, MTAS (Putwain et al., 2020; von der Embse et al., 2021) and the evaluation of the TAQ C, Test Anxiety Questionnaire for Children (Donolato et al., 2019).

Finally, the same process of measurement invariance was undertaken for the Self-efficacy in Test Taking data as for the Children's Test Anxiety Data, to test the confirmed subscales across timepoint, intervention and control and gender groupings.

7.1.1.1 A Word About Fit Indices. The fit indices reported here are a combination of absolute fit indices (*Chi Square* and *SRMR, standardized root mean square residual*), and parsimony correction indices (*RMSEA, root mean square error of approximation*) and the comparative fit indices of *CFI (Comparative Fit Index)* and *TLI (Tucker-Lewis Index).*

The absolute fit index of *Chi Square* is known to be compromised and can be inflated by sample size. It is always recommended to report *Chi Square* but alongside other fit indices that can be relied upon more heavily (Brown, 2015).

SRMR (standardized root mean square residual) is the measure of the difference in correlations which are input from the sample matrix to the correlations in the model matrix. The closer to 0 the better the model fit (Brown, 2015). Hu and Bentler (1999) suggest .08 as a good guideline. All of the tested models for the Adapted Thoughts' subscale return a value of 0.04 to 0.07 suggesting a good fit. (see Table 7.14; Table 7.15)

The *RMSEA* (*root mean square error of approximation)* parsimony correction value is an error of approximation measure and measures how well the model fits the population. Hu and Bentler (1999) suggest values of 0.06 or below.

Browne and Cudeck (1992) developed the Close fit measure of *pCLOSE*. This measure is the statistically significant test for *RMSEA*. *pCLOSE* is reported in this study and a value above the significance value of .00 is sought. However, a non-significant *pCLOSE* should be interpreted alongside the overall model fit (Brown, 2015).

The new Adapted Thoughts' subscale that was tested for the data in this study and, in comparison to the original thoughts' subscale for the data in this study has *pCLOSE* values that are all non-significant indicating good model fit (see Table 7.15).

Comparative fit indices of *CFI* and *TLI* evaluate the fit of the sample specified model to a baseline model. Brown (2015) recommends using *CFI* and *TLI* as they are well-behaved comparative fit indices. Their values should be as near to 0.95 or greater. In the confirmatory factor analysis for the original thoughts' subscale, none of the comparative fit indices in any of the groups or timepoints reaches 0.90 (see Table 7.9) whereas in the confirmatory factor analysis for the Adapted Thoughts' subscales, tested in twelve different conditions, the *CFI* and *TLI* are > .90 in all but one case (see Table 7.14). These results suggest that the Adapted Thoughts' subscale is a better measure than the original thoughts' subscale for the data in this study.

7.1.1.2 Issues for Likert Scale Data. The descriptive statistics and the histograms of responses to the items on the questionnaires for the different groups and timepoints show that the data are not normally distributed. This result is a common state that is assigned to Likert type data. (Norman, 2010)

Likert (1932) did not intend for each item of a questionnaire to be described separately. Likert suggested that the Likert scale (the original is a five-point scale) would generate an overall feel for the psychological construct of attitude which is what the original Likert scale was designed to measure. Rather than analysing the individual items, Likert combined the individual items via summation or taking the arithmetical mean. Likert data are treated as interval data in this study.

Boone and Boone (2012) discuss how to analyse Likert data. In this brief article, the rationale for treating Likert data as interval data is outlined. Although there are Likert items on the questionnaires, that should merit treating the data as ordinal, no individual questions are analysed. It is only sum scores on the Adapted Thoughts' subscale and sum scores on the Self-efficacy in Test Taking subscales that are analysed in this study and as such, the data can be treated as interval data and the corresponding parametric tests applied. Histograms for response frequencies and itemised descriptive statistics have also been included for detail (see Tables G.1 -G.8; H.1 -H.4; Figures G.1 -G.8; Figures H.1-H.14)

Sullivan and Artino (2013), highlight the importance of being able to interpret Likert data accurately in the fields of medicine and education. They refer to the work of Dr. Geoff Norman. Norman (2010) reviewed the controversy surrounding whether or not Likert data can be treated as interval data and the conclusion is that using parametric tests is more robust for Likert data. Parametric tests are more likely to yield the right answer even when statistical assumptions, such as normal distribution of data, are not met.

Furthermore, Carifio and Perla (2007) in listing ten myths about Likert data propose that whether Likert items are interval or ordinal, is not important as Likert scale data, can be taken to be interval. The authors completely dispel the myth that only non-parametric tests should be used for Likert data saying that "*F* is not made of glass". (Carifio & Perla, p.110)

This study uses analyses of the sum scores of subscales for the Adapted Thoughts' subscale from the Children's Test Anxiety Scale data, and the Self-efficacy in Test Taking and Little Self-efficacy Scale in Test Taking subscales from the Self-efficacy data.

7.1.1.3 Justification for Factor Analysis. A brief history of the Children's Test Anxiety Scale instrument will give the reader the context for the decision to conduct exploratory factor analysis after an initial confirmatory factor analysis, on the known measure of the Children's Test Anxiety Scale (Wren & Benson, 2004).

The original Children's Test Anxiety Scale (Wren & Benson, 2004) questionnaire was developed after trying several models and the initial 50 item questionnaire was reduced to 30 items and then to 25. In this study, the 30 items scale was used as per the recommendations in the original paper (Wren & Benson, 2004). The authors stated that the 25 items model did not improve the reliability significantly from the 30 items model. Interestingly, the four questions which show skew or kurtosis, or both, in the descriptive statistics of this study: Q8 'face feels hot', Q17 'head hurts', Q23 'hand-shakes' and Q25 'go to the bathroom' all belong to the *autonomic* subscale. Of these, Q8 and Q25 were removed to create the 25 items version of the questionnaire in the original development of this scale.

Wren and Benson (2004) walk the reader through the process of establishing the 50 items scale validity and the use of *alpha* coefficients to test the reliability of the overall scale and the three subscales. Confirmatory factor analysis of the 30 items scale, following the removal of 20 items, is reported in fit indices statistics for the initial sample and a cross validation sample. The authors report that the individual item loadings varied somewhat across the two samples. Nine of the 30 items had loadings that differed by > 0.10. The authors also wondered whether the correlations between thoughts and the autonomic reactions' subscale were too high at 0.81 in the validation sample. Multicollinearity is measured at $r = > .90$ (Field, 2012)

I decided to explore the factor structure of the Children's Test Anxiety Scale for three reasons: 1. The uncertainty around item loadings; 2. Developments in the test anxiety construct indicating a need to consider test anxiety as a more multidimensional construct (Putwain et al., 2020) 3. The high *RMSEA* values from my initial confirmatory factor analysis (see Table 7.3)

In commenting on the internal construct of the validity of the instrument, having tested the instrument across a validation sample, Wren & Benson (2004) say "in comparing specific loadings across two samples, it was noted that 9 of the 30 items had loadings that differed by > 0.10 (six of the nine were lower in the validation sample), thus while the three-factor structure appears stable over the two samples, individual item loadings varied somewhat. Furthermore, the correlation among the Thoughts and Autonomic Reactions factors is perhaps too highly correlated in the validation sample (0.81) to say the factors are distinct". (Wren & Benson, 2004, p. 237-238)

Given the variability of the factor loadings encountered by the original scale

validation, I felt it was appropriate to proceed with an exploratory factor analysis of the Children's test Anxiety Scale, as it is critical to have the knowledge of loadings between items and factors before conducting confirmatory factor analysis. (Worthington & Whittaker, 2006).

Moreover, since the Children's Test Anxiety Scale (Wren & Benson, 2004) was created, test anxiety theory has developed. von der Embse et al. (2021) discuss the issue of new measures for test anxiety needing to represent the multidimensional nature of the test anxiety construct and point out that the most commonly used test anxiety measure is 40 years old.

7.1.2 Evaluating the Intervention

Once the items on the subscales for both questionnaires are determined from the factor analyses, further analyses on the sum scores obtained from these subscales are used to evaluate the effect of the intervention and answer the research questions. These analyses involve using 2 x 2 mixed ANOVAs and where there is an interaction, a *Tukey Honest Significant Difference* post hoc test. Homogeneity of slopes was tested prior to an ANCOVA to check the role of the pre-test score as a covariate on the data and where this assumption was not met, multi-level linear modelling was used. These are the main analyses for Research Questions 1, 2a and 3; the questions that evaluate the intervention. Chapters 9 discusses the results, interpretation and implementation of the intervention group and Chapter 10 reflects on the intervention and evaluates it further.

7.2 Part 1, The Children's Test Anxiety Scale Data

The data from the 30-item questionnaire were reduced through factor analyses. Following the initial confirmatory factor analysis where the fit statistics of the original thoughts' subscale were unacceptable and conscious of theoretical developments in the test anxiety construct, I used exploratory factor analysis. In this process, I was guided in testing the number of factors by the scree test (Cattell, 1961) and the parallel test (Horn, 1965) as recommended by Brown, (2015) that identifies the number of factors through eigenvalues (see Table 7.11) according to the Kaiser-Guttman rule (Kaiser, 1970; Guttman, 1954)

Confirmatory factor analysis followed the exploratory factor analysis to establish the best subscale with which to measure the worry element of test anxiety in my data. In determining the subscale to use, I have looked at *RMSEA, CFI* and *TLI* values.

7.2.1 Descriptive Statistics for the Individual Items

The following statistical analyses were conducted: mean, standard deviation, skew, kurtosis, and standard error for all of the items on the Children's Test Anxiety Scale at pre-test and post-test (timepoint sets) and intervention and control (groups sets) and for the original thoughts' subscale, as identified by Wren & Benson (2004). (see Appendix G, Table G.1 – G.8)

The descriptive statistics highlighted some items, all of which belong to the original Wren and Benson, (2004) autonomic reactions' subscale of the Children's Test Anxiety Scale questionnaire, with skew and /or kurtosis. The values for skew and kurtosis for normal distribution were taken from Field, 2009. (Skew + /-2, kurtosis +/-2). If we take the values of -2 to + 2, there was no skew or kurtosis in the data with the pre-test questionnaire with the intervention group. However, a pattern of skew and kurtosis emerges with some items in the intervention post-test group and in the control group in pre-test and post-test conditions. (see Table 7.8)

Table 7.8

Skew and Kurtosis Values for Some Items on the Children's Test Anxiety Scale Data, n = 85

Question	Timepoint	Group	Skew	Kurtosis
Q8	Post-test	Intervention		2.02
Q17	Pre-test	Control		2.01
Q17	Post-test	Intervention		3.03
Q17	Post-test	Control		2.12
Q23	Post-test	Intervention		3.14
Q25	Pre-test	Control	2.52	5.54
Q25	Post-test	Intervention	2.46	7.27
Q25	Post-test	Control	1.85	2.46

Note. Acceptable values of Skew + 2, − 2. Acceptable values of Kurtosis + 2, − 2

If the majority of the subsequent analyses were to be carried out with the whole 30 item measure, it might be worth omitting Q25 'While I am taking tests, I have to go to the bathroom' as its kurtosis values are repeatedly high.

Histograms taken for both timepoints pre and post, and both groups, intervention, and control, for the original subscale on the questionnaire, show the frequency distributions for the responses (see Appendix G, Figures G.1- G.4).

The histograms illustrate that the data is not normally distributed. However, it is common for self-report items to have non-normal distribution and the decision to use parametric rather than non-parametric tests in this study is an informed one. Justification for this decision is found in section 7.1.1.2.

Violin plots show the frequency of responses for the Adapted Thoughts' subscale at pre-test and post-test timepoint and for the intervention and control groups. (see Appendix G, Figures G.5 - G.8)

7.2.2 Factor Analysis

Following the justification for conducting factor analysis on a known measure, this section explores the assumptions for exploratory and confirmatory factor analyses for the Children's Test Anxiety Scale data and reports the goodness and badness of fit statistics. The section finishes by detailing the items to be included on an Adapted Thoughts' subscale that is used to measure the effect of the intervention. For all tables of results that are not included in the body of the thesis (see Appendix G)

7.2.2.1 Initial Confirmatory Factor Analysis. The assumptions for confirmatory factor analysis are multivariate normality, a sufficient sample size (see 7.2.3.2), random sampling (see comment in Data Analytic Plan) and an a priori model specification (Brown, 2015). The descriptive statistics (see Appendix G, Tables G.1 - G.8) have already shown that there was a pattern of skew and kurtosis with some of the items from the questionnaire, in the data. Shapiro-Wilk test for the pre and post timepoints and the intervention and control groups show clearly that the data is not normally distributed (see Appendix G, Table G.9)

The data do not have a normal distribution as is typical of Likert type data. Therefore, in the final confirmatory factor analysis process, an "MLM" estimator with Satorra-Bentler method using Lavaan© in RStudio© was used in the maximum likelihood extraction. This estimator is recommended by Brown (2015) to avoid incorrect parameter estimates.

Reasons for using confirmatory factor analysis and subsequent measurement invariance on a smaller sample size are because the premise of one size does not fit all and this premise is applied to this study (see 7.2.3.1.).

The assumption of an a priori model was met through the original Children's Test Anxiety Scale (Wren & Benson, 2004) and through fear appeals research for primary (elementary) aged students (Putwain & Best, 2011) and a translation of the Children's Test Anxiety Scale (Aydun & Bulgan, 2017).

Confirmatory factor analysis was run on the data of the original 13 item thoughts' subscale (see Table 7.9) to test the fit indices of the model to the data. Hu and Bentler (1999) suggest the following values as a guide for a reasonably good fit using the ML, maximum likelihood method of extraction: *Chi- square* – the lower the better; *RMSEA* close to 0.06; *PClose* > 0.05; *CFI* and *TLI* close to 0.95 or greater.

Table 7.9

Fit Indices Statistics for the Original Thoughts' Subscale with the Children's Test Anxiety Scale Data in This Study, n = 85, N = 340.

Timepoint and Group	X^2	*df*	*p*	*RMSEA*	*pCLOSE*	*CFI*	*TLI*	*SRMR*
Pre -test I	118.54	65	.00	0.10	.00	0.85	0.82	0.08
Post-Test I	134.39	65	.00	0.11	.00	0.87	0.85	0.08
Pre-test C	163.54	65	.00	0.13	.00	0.76	0.71	0.10
Post-test C	176.89	65	.00	0.14	.00	0.80	0.76	0.08
All data	220.55	54	.00	1.00	.00	0.89	0.87	0.06

Note. I = intervention group. C = Control group

Note. X^2 = *Chi-square; df = degrees of freedom*; *RMSEA = Root Mean Square Error of Approximation*; *pCLOSE* = value of *RMSEA*; *CFI = Comparative Fit Index*; *TLI = Tucker-Lewis Index*; *SRMR = Standardized Root Mean Square Residual*

Across all the data and also for intervention and control groups, at pre and post timepoints, the fit indices suggest that the original thoughts' subscale is not a good fit for the

data in this study. All of the *RMSEA* measures were 0.10 or more. For *SRMR*, the measure was 0.08 apart from the last condition and neither *CFI* nor *TLI* were above 0.90 for any of the groups or timepoints or with the whole data set $N = 340$.

Cronbach's *alpha* was also measured for this subscale with this data in the different conditions. (see Table 7.10)

Table 7.10

Cronbach's Alpha for the Original Thoughts' Subscale, n = 85, N = 340

Timepoint	Group	*Cronbach's Alpha*
Pre-test	intervention	0.80
	control	0.75
Post-test	intervention	0.91
	control	0.75
All Data		0.88

Note. Children's Test Anxiety Scale Data Across the Intervention and Control Groups at Pre and Post Timepoints.

Although the *alpha* values are acceptable (e.g., Field & Hole, 2003), confirmatory factor analysis is used to estimate the scale reliability in a way that avoids the problems associated with the more traditional methods such as Cronbach's *alpha*. (Brown, 2015). Exploratory factor analysis should show what is happening to the loadings onto the latent variables or factors that have resulted in the unacceptable *RMSEA* values (see Tables 7.12; 7.13).

7.2.2.2 Follow Up Exploratory Factor Analysis. The first assumption to meet for exploratory factor analysis is the Kaiser-Meyer-Olkin test for Sampling Adequacy, (Kaiser, 1970). It can be calculated for individual and multiple variables and is the ratio of the squared correlation between the variables to the squared partial correlation between variables. (Field, 2012). The value varies between 1 and 0 and are detailed in flamboyant terms by Kaiser himself:

in the .90s marvellous,

in the .80s meritorious,

in the .70s middling,

in the .60's mediocre,

in the .50's miserable,

below .50 unacceptable.

The Kaiser-Meyer-Olkin sampling values for the questionnaires involved in this study with both the intervention groups and the control groups for the Children's Test Anxiety Scale at the pre and post timepoints indicate that the data would benefit from factor analysis. All but one of the results are in the upper section of Kaiser's measure. (see Appendix G, Table G.10)

A second assumption to meet for exploratory factor analysis is Bartlett's Test of Sphericity. This test examines whether a variance-covariance matrix is proportional to an identity matrix. (Field, 2012). This test establishes if there is a redundancy between any of the variables that we can summarise with a number of factors. The null hypothesis of the Bartlett's Test for Sphericity is that the variables are orthogonal or not correlated. The test needs to establish if there are significant correlations amongst some of the variables. For factor analysis to work, some relationships between the variables are needed. A significant test of Bartlett's Test for Sphericity is required.

The Bartlett's Test for Sphericity for the questionnaires involved in this study with both the intervention groups and the control groups for the Children's Test Anxiety Scale at the pre and post timepoints indicate that factor analysis for the Children's Test Anxiety Scale data may be useful because all the results were significant. (see Appendix G, Table G.11)

A third assumption to meet prior to exploratory factor analysis is about the correlations for the Children's Test Anxiety Scale data for both the intervention and control groups and at pre and post timepoints. However, there are assumptions that need to be met before running correlation analyses: multicollinearity and singularity.

Although, it is hoped that some of the 30 items show a relationship to each other, if the relationship is too high $r = > .90$, or are perfectly correlated, singularity, $r = 1$ (Field, 2012) then there may not be identifiable factors. Field (2012) refers to mild multicollinearity not being a problem for factor analysis, but extreme multicollinearity is a problem. Correlation coefficients and significance values were calculated using Corrplot © in

RStudio© to see if any item had the majority of values greater than 0.5. The function 'hclust' was also used to group the variables hierarchically.

Although, it is known from the original Children's Test Anxiety Scale paper (Wren & Benson, 2004) that the items correlate well and without multicollinearity, there was a concern about the correlation between autonomic reactions and thoughts' subscales in the validation sample. A value of *alpha* = 0.81 was recorded. In this study, correlation plots were generated for both groups at both timepoints to get a visual interpretation of the way the variables correlate with each other. Visual reference charts as well as numeric ones are provided (see Appendix G, Figures, G.9 – G.16). The method chosen was Pearson as although the data is not normally distributed it is not a small sample size < 30 (Field, 2012)

Multicollinearity was also tested by calculating the determinant of the *R* matrix ensuring that the value is greater than 0.00001 (Field, 2012). In the case of multicollinearity, Field (2012) recommends looking through the items for variables that correlate very highly ($R > .8$) and to think about eliminating these. He says it is a process of trial and error. Given the determinant values (see Appendix G, Table G.12) and the evidence for multicollinearity, the decision was made not to remove items at this stage but after the exploration of items onto factors. Furthermore, as a single subscale for thoughts is being sought from all of the items, it is not necessary to remove any items by trial and error at the outset of exploratory factor analysis; this removal can be done as the process concludes. In addition, the correlation coefficient values were checked to see that they were in the range of + 1 / − 1. (Field, 2013). (see Appendix G, Figures G.9 – G.16).

The fourth and last assumption to meet prior to exploratory factor analysis is the assumption of normality. This assumption was tested in the first confirmatory factor analysis stage with the Children's Test Anxiety Scale data for pre and post-test timepoints and groups using the Shapiro-Wilk test. (see Appendix G, Table G.9) As is typical of Likert data, the distribution of scores was not normal for any timepoint or group as no significance values were greater than .05.

7.2.2.2.1 Decisions in the Exploratory Factor Analysis Process. Having considered the assumptions for exploratory factor analysis in the Kaiser-Meyer-Olken test for Sampling Adequacy and Bartlett's Test of Sphericity and consulted multicollinearity and singularity of correlation values, it is now necessary to decide on the factor extraction method.

In this study, exploratory factor analysis for all the data and for the intervention and control groups at both pre and post timepoints was conducted using scree tests (Cattell, 1966) where the line of the graph acts as a cut-off point to determine factors. In addition, parallel analysis (Horn, 1965) which works by creating a random data set with the same number of observations and variables to determine factors, and the use of the Kaiser-Guttman rule (Guttman, 1954; Kaiser, 1970) where eigenvalues of > 1 determine how many latent variables or factors there are in a particular data set, was also used. (see Table 7.11)

The sample size for each group and each timepoint is $n = 85$ and a value of .4 (salient) Brown (2015) or more was considered as a loading onto the factor. Grice (2001) refers to the importance of the decision-making processes involved in factor extraction and how small differences in method can affect the results.

The method of extraction was ML, maximum likelihood. This method is the most appropriate given that the observed measures in the Children's Test Anxiety Scale data set do approximate an interval level measurement scale. (Brown, 2015). The main advantage in using maximum likelihood as the method for factor extraction is that it will evaluate how well the factor solution can reproduce the relationships among the indicators in the input data statistically and provides fit indices measures. (Brown, 2015). The Satorra-Bentler "MLM" estimator was used to avoid biased standard errors (Brown, 2015).

It is important to explore and ascertain the optimal number of factors in exploratory factor analysis. Choosing too few or too many factors can result in considerable error (E.g., Costello & Osborne, 2005). Brown (2015) suggests that a firm theoretical background and previous experience with the variables will help in the evaluation of the factors determined.

The initial confirmatory factor analysis, the exploratory factor analysis, and the existence of a priori models have contributed to my feeling of familiarity with the items on the measure. The pilot study and test anxiety theory covered in the literature review, plus the knowledge from other studies that use the Children's Test Anxiety Scale, as a measure for test anxiety (e.g., Putwain & Best, 2011), provide a firm theoretical background.

The rules for sample size in exploratory factor analysis have some flexibility. Costello and Osborne (2005) suggest that sample size should be reached through examining the nature of the data and the strength of the item loadings onto the factors. The sample for the intervention and the control group were $n = 85$ respectively, and the number of items on the

instrument is thirty. I looked at each condition, pre-test / post-test, intervention, and control to gain a picture of how the items were loading.

The factors were examined to avoid cross-loadings or no loadings. Factors need to be clear with several indicators that have strong factor scores aligned to it. Factors that have two or three indicators only are not strong Brown, (2015)

The rotation method selected was oblique. Rotation is performed to give the solution with the best simple structure (Thurstone, 1947). Oblique rotation allows the factors to correlate. Orthogonal rotations produce factors that are uncorrelated (Costello & Osborne, 2005) and given that Wren and Benson (2004) report a high correlation between the autonomic reactions and thoughts' subscales, oblique rotation using the 'promax' function in RStudio ©is the best choice for this study. If there is a suspected correlation between the factors, then the oblique rotation will give a more accurate representation of these relationships. (Brown, 2015)

I am satisfied that I have fulfilled the criteria suggested by Brown (2015) for the exploratory factor process. I have chosen the correct estimator, determined the appropriate number of factors through scree plot and parallel analysis, used the correct rotation given what I know of the measure, interpreted the factors, and evaluated the quality of the solutions through eliminating poorly defined factors and poorly behaved items and by testing various three and four factor models. I have not been able to test the factor structure on another sample. I have not created factor scores to check for indeterminacy as I will be following the exploratory factor analysis with confirmatory factor analysis where indeterminacy is not such an issue (Brown, 2015).

Table 7.11

Eigenvalues >1 for the Children's Test Anxiety Scale Data, n = 85, N = 340

Timepoint	Group	Scree Test	Parallel Analysis
Pre-test	intervention	3 factors, eigenvalue > 1	3 factors, eigenvalue > 1
Pre-test	control	3 factors, eigenvalue > 1 (one further factor just on the line)	4 factors, eigenvalue >1
Post-test	intervention	3 factors, eigenvalue > 1	3 factors, eigenvalue > 1 (one further factor just on the line)
Post-test	control	2 factors, eigenvalue > 1 (two further factors on the line)	3 factors, eigenvalue > 1 (one further factor just on the line)
All data		3 factors eigenvalue > 1	3 factors eigenvalue > 1

Note. Scree test (Cattell, 1966), Parallel analysis (Horn, 1965)

Given the results of the scree test (Cattell, 1966) and the parallel analysis (Horn, 1965), the decision was made to explore a three factor and a four-factor model for all the groups at the different timepoints.

For ease of reference, I have included fit statistics, *chi-square*, *degrees of freedom* and the *p* value of *chi square* for both the intervention and control groups and at the pre and post timepoints in the body of the thesis (see Table 7.12). For factor loadings for the three factor and four factor models (see Table 7.13).

Table 7.12

Fit Indices Statistics for Exploratory Factor Analysis Using Lavaan© in RStudio©, n = 85.

Group, Timepoint, Factors	*RMSEA*	*RMSR*	*TLI*	X^2 / *df*	*p*. Value of X^2	*alpha*
Intervention Pre-test x3	0.04	0.06	0.92	394.12 /348*df*	.04	0.88
Intervention Pre-test x4	0.03	0.06	0.95	348.58 /321*df*	.139	0.87
Intervention Post-test x3	0.07	0.07	0.80	487.56 /348*df*	< .001	0.92
Intervention Post-test x4	0.06	0.06	0.83	429.57/321*df*	< .001	0.91
Control Pre-test x3	0.06	0.08	0.77	471.15 /348*df*	< .001	0.84
Control Pre-test x4	0.05	0.07	0.83	402.13 /321*df*	.0014	0.84
Control Post-test x3	0.06	0.07	0.84	459.30/348*df*	< .001	0.87
Control Post-test x4	0.05	0.06	0.88	396.55 /321*df*	.002	0.88

Note. Children's Test Anxiety Scale Intervention and Control Groups at Both Pre and Post-test and with Three and Four Factor Extraction and with Cronbach's *Alpha*

Note. RMSEA = Root Mean Square Error of Approximation; *RMSR = Root Mean Square Residual*; *TLI = Tucker-Lewis Index*; X^2 *= Chi-square; alpha = Cronbach's alpha.*

Table 7.13

Factor Loadings for the Different Conditions - Children's Test Anxiety Scale for This Study

Item	Condition							
	1	2	3	4	5	6	7	8
Q1	0.53[t]	0.42[t]	0.54[t]	0.45[t]		0.47[4]		0.47[t]
Q2	0.65[a]	0.73[a]	0.55[a]	0.59[a]	0.61[a]	0.58[a]	0.88[a]	0.73[a]
Q3	0.80[o]	0.74[t]	0.69[o]	0.60[o]		0.90[o]		0.74[4]
Q4	0.40[t]		0.56[a]		0.40[a]	0.60[a]	0.84[a]	0.60[a]
Q5	0.74[t]	0.71[t]	0.53[t]	0.58[t]	1.02[t]	1.03[t]	1.05[t]	1.12[t]
Q6	0.52[o]				0.62[t]	0.57[t]	0.45[t]	0.46[t]
Q7	0.52[o]	0.92[o]	0.41[o]	0.74[4]	0.52[o]		0.66[o]	0.53[o]
Q8	0.80[a]	0.84[a]	0.65[a]	0.84[a]	0.44[a]	0.49[a]	0.46[a]	0.69[a]
Q9	0.86[t]	0.92[t]	0.84[t]	0.92[t]	0.70[t]	0.65[t]	0.85[t]	0.91[t]
Q10	0.72[a]	0.64[a]	0.54[a]	0.56[a]	0.65[a]	0.77[a]	0.50[a]	0.62[a]
Q11	0.41[t]	0.42[t]	0.71[t]	0.73[t]			0.49[t]	0.55[t]
Q12				0.41[a]	0.58[o]	0.55[o]	0.74[o]	0.71[o]
Q13			0.49[t]	0.43[o]	0.77[o]	0.79[4]		
Q14	0.44[t]	0.45[o]	0.56[o]	0.41[o]		0.60[o]	0.69[o]	0.57[o]
Q15	0.43[o]	0.56[4]	0.53[t]	0.66[o]	0.52[o]			0.41[t]
Q16			0.53[t]	0.44[t]	0.44[t]	0.40[t]		0.49[t]
Q17	0.53[a]	0.45[a]	0.49[a]		0.60[a]	0.60[a]		0.49[a]
Q18	0.71[o]	0.47[o]	0.60[o]	0.45[o]		0.64[o]		0.77[4]
Q19	0.63[t]	0.57[t]	0.65[t]	0.75[t]	0.76[t]	0.69[t]	0.69[t]	0.76[t]
Q20	0.76[a]	0.78[a]	0.57[a]	0.77[a]	0.45[o]		0.64[a]	0.87[a]
Q21	0.43[a]	0.41[a]	0.56[t]	0.51[t]	0.47[o]	0.56[a]	0.44[o]	
Q22	0.42[o]				0.47[o]		0.85[o]	0.75[o]
Q23	0.69[a]	0.59[a]	0.69[a]	0.49[a]	0.53[a]	0.53[a]	0.62[a]	0.63[a]
Q24	0.64[t]	0.78[t]	0.72[t]	0.66[t]				
Q25	0.49[o]		0.49[a]	0.44[4]				
Q26		0.56[o]	0.40[o]	0.48[4]	0.74[o]	0.44[4]		
Q27	0.80[t]	0.83[t]	0.78[t]	0.89[t]	0.50[t]	0.46[t]	0.49[t]	0.43[t]
Q28	0.47[a]	0.48[t]	0.63[t]	0.74[t]	0.46[a]	0.61[a]	0.44[a]	
Q29	0.47[a]	0.48[t]	0.69[t]	0.73[t]	0.40[o]		0.41[o]	
Q30	0.90[o]	0.66[o]	0.68[o]	0.68[o]	0.56[o]		0.49[o]	0.43[o]

Note: [a] = autonomic subscale, [t] = thoughts' subscale, [o] = off task behaviours subscale,

[4] = fourth unnamed factor. For conditions see following list:

Note. The different three and four factor intervention and control groups at pre and post timepoint are detailed in the conditions below.

Condition 1*: Loadings for Three Factors with the CTAS Pre-test Intervention Group*

Condition 2*: Loadings for Four Factors with the CTAS Pre-test Intervention Group*

Condition 3: *Loadings for Three Factors with the CTAS Post-test Intervention Group*

Condition 4: *Loadings for Four Factors with the CTAS Post-test Intervention Group*

Condition 5: *Loadings for Three Factors with the CTAS Pre-test Control Group*

Condition 6*: Loadings for Four Factors with the CTAS Pre-test Control Group*

Condition 7: *Loadings for Three Factors with the CTAS Post-test Control Group*

Condition 8: *Loadings for Four Factors with the CTAS Post-test Control Group*

7.2.2.3 Trial Adaptations for confirmatory factor analysis of the subscale adapted thoughts. The loadings were examined in all the different conditions for cross loadings, no loadings, anomalous loadings and consistency of loadings to factors (see Table 7.8). From these results, three different adaptations were tried with further confirmatory factor analysis.

Adaptation 1: Thoughts: Q1 'Wonder if I will pass', Q4, 'Feel nervous', Q5 'Think I will get a bad result', Q9 'I worry about failing', Q11 'I worry about doing something wrong', Q16 'I should have studied more', Q19 'I think most of my answers are wrong, Q27 'I think how badly I am doing', Q28 'I feel scared'. (Q4 and Q28 incorporated from the autonomic reactions' subscale)

Adaptation 2: Thoughts: Q1 'Wonder if I will pass', Q4, 'Feel nervous', Q5 'Think I will get a bad result', Q9 'I worry about failing', Q11 'I worry about doing something wrong', Q16 'I should have studied more', Q19 'I think most of my answers are wrong, Q27 'I think how badly I am doing' (Q4 incorporated from the autonomic reactions' subscale).

Adaptation 3: Thoughts: Q1 'Wonder if I will pass', Q5 'Think I will get a bad result', Q9 'I worry about failing', Q11 'I worry about doing something wrong', Q16 'I should have studied more', Q19 'I think most of my answers are wrong, Q27 'I think how badly I am doing', Q28 'I feel scared'. (Q28 incorporated from the autonomic reactions' subscale)

Wren & Benson (2004) and the development of the Children's Test Anxiety Scale note loadings being inconsistent across the validation sample and this inconsistency is

mirrored in the exploratory factor analysis findings from this study with items Q4 ‘I feel nervous’ and Q28 ‘I feel scared’ cross loading onto the autonomic reactions and thoughts’ subscales.

The items that are omitted from the Adapted Thoughts’ subscale and the reasons for the decisions are detailed here: Q6 (hard to remember the answers) failed to load in three conditions, loaded onto thoughts’ three times and off task behaviours twice; Q13 (think what my result will be) failed to load in four conditions, loaded onto thoughts’ once, two loadings onto off task behaviours and one loading onto a fourth factor; Q15 (wonder if my answers are right) failed to load in two conditions, three loadings onto off task behaviours, two loadings onto thoughts’ and one loading onto a fourth factor; Q21(worry how hard the test is) failed to load in one condition, three loadings onto autonomic reactions, two loadings onto off task behaviours and two loadings onto thoughts; Q24 (think what will happen if I fail) failed to load in four conditions on any subscale and loaded onto the thoughts’ subscale for all four intervention conditions; Q29 (worry what my parents will say) failed to load in two conditions, two loadings onto off task behaviours, one loading onto autonomic reactions and three loadings onto the thoughts’ subscale.

In the context of scale development research, Worthington and Whittaker’s (2006) systematic review of new scale development articles from 1995-2004, found ten papers where confirmatory factor analysis followed directly on from exploratory factor analysis in confirming the validity of a scale measurement. This method is supported also by Fabrigar et al. (1999), who refer to the process as being useful in conjunction with one another. Others suggest splitting the data set in half and using exploratory factor analysis on half of the data and confirmatory factor analysis on the other half. (e.g., Putwain et al., 2020). This process necessarily depends upon sample size (Fabrigar et al.,). Gerbing and Hamilton (1996) suggest that exploratory factor analysis prior to cross-validation with confirmatory factor analysis is useful. I did not split the data in half in this study and as such the Adapted Thoughts’ subscale needs to be cross validated with another sample.

Following the exploratory factor analysis where the emphasis was on exploring the correlation coefficients to determine the latent factors to establish the subscales, a further confirmatory factor analysis was conducted on the Children’s Test Anxiety Scale, adapted version of the thoughts’ subscale.

7.2.2.4 Confirmatory Factor Analysis of the Adapted Thoughts' Subscale. The benefits of confirmatory factor analysis to exploratory factor analysis are well documented (Brown, 2015). The process of confirmatory factor analysis in this study is driven by the prior research that confirmed the three factor or subscale structure of the Children's Test Anxiety Scale instrument for measuring test anxiety (Wren & Benson, 2004).

Whilst exploratory factor analysis allows the variables to load freely onto the factors, in confirmatory factor analysis all aspects of the model to be tested are specified. Measurement invariance measures a latent variable across different groups and timepoints to test for consistency of the subscale with different groups and at different times. In this study, measurement invariance is carried out for the pre and post-test timepoints and with the intervention and control groups and with gender groups to test for consistency in the subscale across different groups.

7.2.2.5 Assumptions for confirmatory factor analysis. As outlined in section 7.2.2.1, there are some assumptions that I had to test on the data before I could run the final analyses.

The different versions of the Adapted Thoughts' subscale were tested across the pre and post timepoints and with the intervention and control groups and with all of the data to see if it is a good measure for the cognitive element of test anxiety or worry in 10-to 11-year-olds. For confirmatory factor analysis fit indices' results for the different adaptations of this subscale (see Table 7.14; Table 7.15)

Overall, the Adapted Thoughts' subscale for the Children's Test Anxiety Scale data is a better fit (see Table 7.14; Table 7.15) than the original 13 item model from 2004 (see Table 7.9)

Table 7.14

Fit Indices Statistics for the Three Different Versions of the Adapted Thoughts' Subscale with the Children's Test Anxiety Scale Data (n = 85)

Adaptation Timepoint Group	X^2	*df*	*p*	*RMSEA*	*pCLOSE*	*CFI*	*TLI*	*SRMR*
Adaptation 1 (Q4+Q28)								
Pre-test I	47.33	27	.09	0.94	0.06	0.92	0.89	0.07
Post-test I	35.55	27	.12	0.06	0.35	0.97	0.96	0.05
Pre-test C	39.94	27	.05	0.75	0.20	0.94	0.92	0.06
Post-test C	49.42	27	.00	0.10	0.04	0.92	0.90	0.07
Adaptation 2 (Q4)								
Pre-test I	24.28	20	.23	0.05	0.46	0.98	0.97	0.06
Post-test I	31.18	20	.05	0.08	0.17	0.96	0.94	0.05
Pre-test C	31.41	20	.05	0.08	0.17	0.92	0.94	0.06
Post-test C	30.10	20	.07	0.08	0.21	0.96	0.95	0.05
Adaptation 3 (Q28)								
Pre-test I	32.91	20	.03	0.09	0.13	0.94	0.91	0.06
Post-test I	26.67	20	.14	0.06	0.34	0.98	0.97	0.05
Pre-test C	22.21	20	.33	0.04	0.57	0.99	0.98	0.05
Post-test C	34.06	20	.03	0.09	0.10	0.95	0.93	0.06

Note. I = intervention group. C = Control group.

Note. X^2 = *Chi-square; df = degrees of freedom; RMSEA = Root Mean Square Error of Approximation; pCLOSE* = value of *RMSEA; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; SRMR = Standardized Root Mean Square Residual*

Table 7.15

Fit Indices Statistics for the Three Different Versions of the Adapted Thoughts' Subscale with the Children's Test Anxiety Scale Data (N = 340)

Adaptation	X^2	*df*	*p*	*RMSEA*	*pCLOSE*	*CFI*	*TLI*	*SRMR*
Adaptation1(Q4, Q28)	97.07	27	.00	0.09	.00	0.93	0.91	0.05
Adaptation 2 (Q4)	52.09	20	.00	0.07	.08	0.97	0.95	0.04
Adaptation 3 (Q28)	62.49	20	.00	0.08	.02	0.96	0.94	0.04

Note. *X^2 = Chi-square; df = degrees of freedom; RMSEA = Root Mean Square Error of Approximation; pCLOSE =* value of *RMSEA; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; SRMR = Standardized Root Mean Square Residual*

Based on the fit indices results from the confirmatory factor analysis and from what is known of the loadings from exploratory factor analysis and taking into account test anxiety theory, the Adapted Thoughts' subscale to be used in the main analyses is Adaptation 3:

Q1 'While I am taking tests, I wonder if I will pass'

Q5 'While I am taking tests, I think I'm going to get a bad result'

Q9 'While I am taking tests, I worry about failing'

Q11 'While I am taking tests, I worry about doing something wrong'

Q16 'While I am taking tests, I think I should have studied more'

Q19 'While I am taking tests, I think most of my answers are wrong'

Q27 'While I am taking tests, I think how badly I am doing'

Q28 'While I am taking tests, I feel scared'.

Although the goodness of fit statistics across the whole data set are slightly better for adaptation 2, the loadings for Question 4 'I feel nervous' are mainly on the autonomic reactions' subscale in the exploratory factor analysis in comparison to Question 28 which loads three times onto the thoughts' subscale and with high values. (see Table 7.13) In

addition, given the similarity in meaning of ‘nervous’ and ‘scared’, I think it is better to use Q28 that loaded onto the thoughts or cognitive subscale.

7.2.3 Measurement Invariance

The Adapted Thoughts’ subscale is adaptation 3 and includes Q28 ‘I feel scared’.

Adaptation 3: Thoughts: Q1 ‘Wonder if I will pass’, Q5 ‘Think I will get a bad result’, Q9 ‘I worry about failing’, Q11 ‘I worry about doing something wrong’, Q16 ‘I should have studied more’, Q19 ‘I think most of my answers are wrong, Q27 ‘I think how badly I am doing’, Q28 (Q28 incorporated from the autonomic reactions’ subscale).

To establish that the Adapted Thoughts’ subscale has the same meaning to groups across the different measures of time and intervention and control groups, multi group confirmatory factor analysis to test for measurement invariance (Finch and French, 2015) has been calculated. The Adapted Thoughts’ subscale has been constructed following an exploratory factor analysis that showed some cross loadings of items onto the thoughts and autonomic reactions. The main difference being tested is item Q28 ‘I feel scared’ as a cognitive element rather than an autonomic reaction.

The principle of measurement invariance is important in establishing scale development (Brown, 2015; von der Embse et al., 2021) and the purpose is to establish an equivalence of meaning for the construct across time and groups. (Putnick & Bornstein, 2016). In this way, the Adapted Thoughts’ subscale will be tested across the intervention and control groups and also across boys and girls for multi group invariance and across timepoints for longitudinal invariance. There are four main steps to testing measurement invariance: configural, metric, scalar and strict.

Configural invariance assumes that the number of latent variables and their observed indicators is the same for all groups of the population. If this result is true, we can say that a model’s parameters are invariant across the groups. Following configural invariance, we test for *metric invariance* where factor loadings are tested across the groups for equivalency. The next step in the measurement model is to test for *scalar* invariance where we examine the invariance of the factor model *intercepts.* The final step in the process is to test for *residual invariance.* This step tests for factor and error invariance. Residual invariance is not a necessary part of the process of testing to be able to test for latent means differences because the residuals are not part of the latent factor. (Putnick & Bornstein, 2016).

The results are reported using *ChiSquare* and its *p* value, although both of these statistics are sensitive to sample size and cannot always be relied upon. Badness of fit index *RMSEA,* goodness of fit *CFI, ΔCFI, TLI* and also *SRMR* are reported.

The results (Appendix G, Tables G. 23 – G. 28) show that measurement invariance holds across the configural to constrained to metric, scalar and strict models for the timepoints. However, this result is not the case with the intervention and control groups in the strict factor to strict error stage where *ΔCFI is* 0.3. Bialosiewicz et al. (2013) suggest that it is unreasonable to expect equality in the strict invariance model across groups or across time. Brown (2015) and Putnick and Bornstein (2016) argue that tests for residual invariances are optional and that ascertaining and evaluating equal error variances is not really necessary in the majority of research studies.

The *ΔCFI* is within the acceptable levels at the configural, metric and scalar stages of the comparison across the intervention and control groups and so this result is accepted as measurement invariance.

Therefore, measurement invariance holds for the Adapted Thoughts' subscale for the Children's Test Anxiety Scale data across the intervention and control groups, the pre-test and post-test timepoints and across the gender groups of boys and girls.

If the configural, metric and scalar models hold invariance across their group or timepoint, it is possible to continue to conduct further tests for population heterogeneity; these are: tests for the equality of factor variances, the equality of factor covariances (not applicable in this study as only one factor is being measured) and the equality of latent means. Brown (2015) suggests that comparison of latent means in confirmatory factor analysis is comparable to, but stronger than, ANOVA. This suggestion is because the comparison of means is being conducted in the context of the measurement model. Latent means and ANOVA were used in this study and are reported in Chapter 8. The results are not compared with each other, but I wanted to use latent means in addition to the more traditional ANOVAs because I had taken the multigroup confirmatory factor analyses to a point where I was able to use the latent means' calculations.

In this study, measurement invariance has been proved across the timepoints of pre and post-test, across gender, and across intervention and control groups. ($N = 170$) This result means that the Adapted Thoughts' subscale is a reliable construct to use with this data and

across the different groups. (see Figure 7.1) However, before this subscale can be used more widely, it needs to be cross validated with another sample.

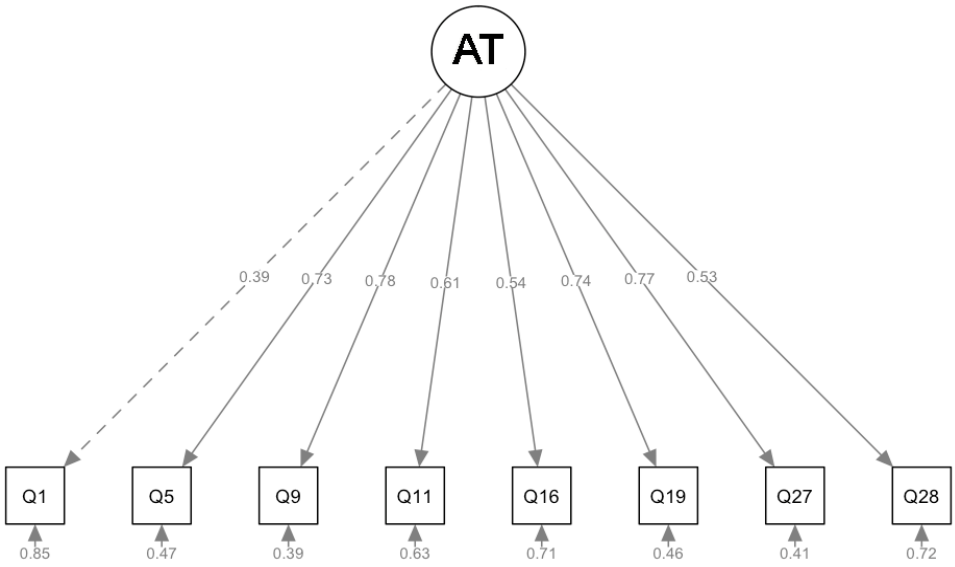
7.2.3.1 Sample Size for Measurement Invariance for the Adapted Thoughts' Subscale, Children's Test Anxiety Scale Data. The required sample size for SEM has been discussed at length over the years. The sample size in this study is $N = 340$ for the intervention and control groups combined with pre-test and post-test for the Children's Test Anxiety Scale data. Marsh et al. (1998) have suggested that when $N = 100$, there should be at least four items per factor in the model. For this part of the study, there are eight items to one factor.

Wolf et al. (2013) debate the issue of sample size requirements and conclude that there is no one size fits all. Various rule of thumb measures evolved over the years, but Wolf et al. (2013) look for a more accurate approach. Using the Monte Carlo method of data simulation techniques, they concluded (amongst other model results) that a one factor, six indicator model requires sample sizes of 90, 60 and 40 with loadings of .50, .65 and .80 respectively. The Adapted Thoughts' subscale has met these sample size requirements.

The estimator used in all the measurement calculations is Satorra Bentler "MLM". This estimator has been used because the data does not have a normal distribution and the Satorra-Bentler method is a robust measure to correct for this broken assumption.

Figure 7.1

The Adapted Thoughts' Subscale for the Children's Test Anxiety Scale Data, N =340



There are eight items with 16 indicators depicted and the loadings or standardized regression coefficients and the error variances. Apart from item Q1, the values are all at 0.50 or more. Cronbach's *alpha* for the Adapted Thoughts' subscale for all the data is *alpha* = 0.82 which is an acceptable level.

7.3 Part II: Self-efficacy In Test Taking

The Self-efficacy in Test Taking instrument was created by me. Putwain and Aveyard (2018) and Putwain and Pescod (2018) refer to control and academic buoyancy in managing the state, trait problem of test anxiety. The aim of the intervention was to make 10-to 11-year-olds feel more self-efficacious about taking tests and thus reduce the worry element of test anxiety. It was necessary to create an instrument to try and measure Self-efficacy in Test Taking as, to my knowledge, there are none.

The Self-efficacy in Test Taking instrument (Appendix A) was designed using Bandura's guidance on constructing self-efficacy questionnaires for different domains (Bandura, 2006). The main theories behind the construction of this questionnaire are test anxiety theory (e.g., Zeidner & Matthews, 2005) and self-efficacy theory (Bandura, 1997). It was important to draw upon the conceptual analysis and expertise of what it takes to succeed in test taking and thus the items on the Self-efficacy in Test Taking instrument reflect the issues of confidence in test taking, a 'can do' attitude towards test taking and represent beliefs about personal abilities in test taking. (Bandura, 2006)

The sample is smaller (n = 59 for the intervention group) and (n = 86 for the control group) than the Children's Test Anxiety Scale data sample (n = 85) for the intervention and control groups respectively. The reason is because I trialled the Self-efficacy in Test Taking questionnaire with my own class, after the pilot and my upgrade viva, before making it part of the main study and was not able to let two schools use the measure until I had trialled it. The Self-efficacy in Test Taking instrument was taken at the same time as the Children's Test Anxiety Scale questionnaire and at pre and post-test timepoints and with both intervention and control groups.

7.3.1 Descriptive Statistics for the Individual Items

Descriptive statistics were calculated for the: mean, standard deviation, skew, kurtosis and standard error for the intervention and control groups and at pre and post-test timepoints (see Appendix H, Tables H.1 – H.4). If we take the accepted value for skew as + 2 / – 2 and kurtosis as + 2 / − 2 as per the values applied to the Children's Test Anxiety Scale data

(Field, 2009), then two questions repeatedly show skew and /or kurtosis in the sample. These are questions five and six. ‘When I take tests, I feel anxious about being in a different room to my usual classroom’ and ‘When I take tests, I feel anxious about the instructor’, respectively. (see Table 7.16)

Table 7.16

Skew and Kurtosis Values of Items 5 and 6 for the Self-efficacy in Test Taking Data, n = 59 for the Intervention Group and n = 86 for the Control Group

Question	Timepoint	Group	Skew	Kurtosis
Q5	Pre-test	Intervention	1.16	−0.10
Q5		Control	1.18	0.34
Q5	Post-test	Intervention	2.02	3.89
Q5		Control	2.10	4.05
Q6	Pre-test	Intervention	0.95	0.02
Q6		Control	1.67	2.59
Q6	Post-test	Intervention	1.66	2.95
Q6		Control	1.38	1.22

Note. Self-efficacy Pre and post-test Intervention and Control Groups. *N* = 145

Acceptable values of Skew + 2, − 2. Acceptable values of Kurtosis + 2, − 2

The Self-efficacy in Test Taking instrument uses a Likert type scale that is similar to the Children’s Test Anxiety Scale instrument. This choice was deliberate in the design of the experiment and made to mirror the Children’s Test Anxiety Scale Likert type scale to make it easier for 10-to 11-year-olds to follow. Likert data is not normally distributed but the skew and kurtosis values of items Q5 and Q6 suggest they should be removed from the questionnaire as they are skewing the data. This process is acceptable practice in initial scale construction. (e.g., Putwain, 2020)

7.3.2 Factor Analysis

The assumptions for the exploratory factor analysis and confirmatory factor analysis are the same as in section 7.2.2.2 (see Appendix H, Tables H.5 – H.9)

The Shapiro-Wilk Test for Normality of Distribution illustrates clearly that the self-efficacy data is not normally distributed. The Shapiro-Wilk test statistic is significant in both the intervention and control groups and at both pre and post timepoints for the self-efficacy data. The data do not have a normal distribution as is typical of Likert type data. Therefore, in the final confirmatory factor analysis process, an "MLM" estimator with Satorra-Bentler method using Lavaan© in RStudio© can be used.

Histograms of the individual items show the response values for all of the self-efficacy data and the skew and kurtosis from Q5 'When I take tests, I feel anxious about the being in a different room to my usual classroom' and Q6 'When I take tests, I feel anxious about the instructor' can be seen clearly and indicate that students rarely feel anxious about the instructor and the room. (see Appendix H, Figure, H.1, Figure, H.2)

Violin plots show the density of responses for the two subscales with item five and item six removed due to the high skew and kurtosis values and as a result of the exploratory and confirmatory factor analyses. (see Appendix H, Figures H.3 – H.10)

Cronbach's *alpha* tests for reliability of measure with the data from the self-efficacy questionnaire were tested in both intervention and control groups and pre and post-test timepoints. These tests were undertaken in R Studio© and the 'check.keys = TRUE' functions applied. This code automatically adjusts reverse worded items to make the questions more uniform in theme. In this case, there are several reverse worded items which R Studio© adjusted for (see Table 7.17)

The items which R Studio© adjusted for are: Q4 'If I find a test hard, I feel I will not pass', Q5 'When I take tests, I feel anxious about being in a different room to my usual classroom', Q6 'When I take tests, I feel anxious about the instructor', Q7 'If I find a test hard, I can't control my negative thoughts about failing', Q8, 'If I find a test hard, I start to doubt my ability to pass' and Q9 'In tests, I panic if I see people finishing before me' (see Table 7.17)

Table 7.17

Cronbach's Alpha for the Self-efficacy in Test Taking Instrument - Intervention and Control Groups, n = 59 intervention, n = 86 control.

Timepoint	Group	*Alpha*	*Alpha*CK	Reverse Coded Items
Pre-test	Intervention	0.44	0.75	Q4, Q5, Q6, Q7, Q8, Q9
	Control	0.28	0.84	Q4, Q5, Q6, Q7, Q8, Q9
Post-test	Intervention	0.41	0.81	Q4, Q5, Q6, Q7, Q8, Q9
	Control	0.46	0.83	Q4, Q5, Q6, Q7, Q8, Q9

*Note. Alpha*CK = Values from the "Check.keys = TRUE" Function in RStudio©

The *alpha* scores suggest that the self-efficacy instrument has internal consistency if the items are reverse coded. If the 'check.keys = FALSE' code is applied, which means the items are not reverse coded, then the *alpha* scores are unacceptable. An acceptable *alpha* score should be 0.7 (Nunally,1978)

The two items with unacceptable skew and kurtosis, Q5 'When I take tests, I feel anxious about being in a different room to my usual classroom' and Q6 'When I take tests, I feel anxious about the instructor' are also reverse coded items. Given that the questionnaire was developed using self-efficacy theory but is brand new in terms of measuring test anxiety self-efficacy, there are no other models to draw upon for comparison and thus exploratory factor analysis is appropriate. However, before completing the exploratory factor analysis, it is necessary to return to the assumptions that the data must meet for exploratory factor analysis.

The first assumption to meet for factor analysis is the Kaiser-Meyer-Olkin test for sampling adequacy and is described at section 7.2.2.2.

In all but one of the groups, the Kaiser-Meyer-Olkin measure for sampling adequacy is in the 'middling' range. This result suggests that the sample is suited for factor analysis.

A second assumption to meet for factor analysis is Bartlett's Test of Sphericity. This assumption is described at section 7.2.2.2.

The results show that *p* is significant in both groups and at both pre and post timepoints. This result suggests the data set is suitable for a reduction technique such as exploratory factor analysis.

A further assumption to meet before exploratory factor analysis is the test for multicollinearity and singularity. This assumption is described at section 7.2.2.2.

The correlation plots show no evidence of multicollinearity or singularity with either the intervention or the control group or at the pre-test and post-test timepoints. (See Figures H.11-H.18).

The self-efficacy instrument and data have met all the assumptions for exploratory factor analysis. For all assumption results (see Appendix H, Tables H.5 – H.9, Figures H.11 – H.18)

7.3.2.1 Exploratory Factor Analysis on the Self-efficacy Instrument Across the Timepoints and with the Intervention and Control Groups. Sample size restrictions for exploratory factor analysis have mainly disappeared. (Costello & Osborne, 2005) and for the self-efficacy data in this study, they are: $N = 290$, $n = 118$ (intervention) and $n = 172$ (control) groups.

Firstly, it is the scree test (Cattell, 1966) that indicates the number of factors for the item loadings. For the self-efficacy instrument, scree tests were run for both the intervention and control groups and at both pre-test and post-test timepoints. In applying the Kaiser-Guttman criterion (Guttman, 1954, Kaiser, 1960) where factors with an eigenvalue > 1.0 should be extracted. In addition, Horn's test of parallel analysis was applied. (see Table 7.18)

Table 7.18

Eigenvalues >1 for The Self-efficacy Data, n =59 (intervention), n = 86 (control), N = 290

Timepoint	Group	Scree Test	Parallel Analysis
Pre-test	intervention	2 factors, eigenvalue > 1	2 factors, eigenvalue > 1
Pre-test	control	1 factor, eigenvalue > 1 and an additional factor is just on the line	2 factors, eigenvalue >1
Post -test	intervention	2 factors, eigenvalue > 1 and an additional factor just on the line	2 factors, eigenvalue > 1 and an additional factor just on the line
Post-test	control	2 factors, eigenvalue > 1	2 factors, eigenvalue > 1
All data		2 factors eigenvalue > 1 (one of these factors is just on the line)	2 factors, eigenvalue > 1 (one of these is just on the line)

Note. Scree Tests (Catell, 1966), Parallel Analysis (Horn, 1965) for the Self-Efficacy Data.

The results of the scree tests (Cattell, 1966) and parallel analyses (Horn, 1965) show that in the pre-test and post-test condition for both the intervention and control groups, the likely number of factors is two, if we use the Kaiser-Guttman criterion. However, the post-test intervention groups suggest a three-factor model. Scree tests (Cattell, 1966) and parallel analysis (Horn, 1965) suggest a two-factor model for the whole data, $N = 290$. Given that one factor sits on the line in several conditions, I decided to run further analyses with one and two factor models.

Underfactoring and overfactoring can compromise the validity of the factor model. (Brown, 2015, Costello & Osborne, 2005). Brown suggests that although exploratory factor analysis is indeed an exploratory process, knowledge of the theoretical background for the instrument and experience with the variables in the instrument is needed to decide on the best number of factors. The theories behind the Self-efficacy in Test Taking scale are self-efficacy (Bandura, 1997); belief in capability in test taking and test anxiety theory (e.g., Spielberger & Vagg, 1995; Zeidner, 1998).

However, the two theories have never been combined before in an instrument to measure self-efficacy in test taking abilities for 10-to 11- year-olds. Therefore, exploratory factor analysis will test the instrument for loadings in both one factor and two factor conditions.

7.3.2.1.1 Decisions Within the Exploratory Factor Analysis Process. The same decisions for process were applied to this data as the Children's Test Anxiety Data as detailed in section 7.2.2.2.1. It is worth recalling the skew and kurtosis of items Q5 'When I take tests, I feel anxious about being in a different room to my usual classroom' and Q6 'When I take tests, I feel anxious about the instructor' from the descriptive statistics. These items were not removed initially from the self-efficacy instrument. However, the loadings of these two items were monitored closely in the exploratory factor analysis process. For the different loadings of the items onto the one and two factor exploratory conditions for both the intervention and control group and at both pre-test and post-test timepoints (see Appendix H, Table H.10) For fit indices for the different conditions and number of factors (see Table 7.19)

Table 7.19

Exploratory Factors Analysis: Fit Statistics for the Self-efficacy Instrument Across Intervention and Control Groups and at Pre-test and Post-test Timepoints for One and Two Factor Extractions.

Condition	*Chi Square*	*df*	*p*	*RMSEA*	*TLI*	*RMSR*
1	256.22	90	< .001	0.12	0.38	0.14
2*	129.50	76	< .001	0.10	0.56	0.10
3	232.79	90	< .001	0.12	0.52	0.14
4*	129.25	76	< .001	0.08	0.75	0.10
5	144.83	90	< .001	0.06	0.87	0.09
6*	67.97	76	< .73	0.00	1.03	0.06
7	198.38	90	< .001	0.09	0.72	0.10
8*	106.64	76	< .01	0.07	0.86	0.08

Note. *Indicates a two- factor extraction. Condition numbers refer to the following conditions:

Intervention Group	Control Group
1= Pre-test x1 factor	5 = Pre-test x1 factor
2 = Pre-test x2 factors	6 = Pre-test x2 factors
3 = Post-test x1 factor	7 = Post-test x1 factor
4 = Post-test x2 factors	8 = Post-test x2 factors

Note. df = degrees of freedom; RMSEA = Root Mean Square Error of Approximation; TLI = Tucker-Lewis Index; RMSR = Root Mean Square Residual

Apart from condition 2, the two factor conditions have more acceptable *RMSEA* values (Hu and Bentler, 1999) and *TLI* values than the one factor conditions. The aim of the exploratory factor analysis is to evaluate the dimensions of the indicators (items on the questionnaire) and to uncover the smallest number of latent variables in the instrument from the data. Loadings have been calculated for the new Self-efficacy in Test Taking instrument over eight different conditions. The results, both loadings onto factors and the goodness and

badness of fit statistics show that a two-factor solution is a better fit than a one factor solution. In determining the subscales to take forward for confirmatory factor analysis, all of the item loadings in all of the conditions were considered and a value of 0.32 as a minimum applied (e.g., Tabachnick & Fidell, 2001).

There is a natural division in the items on the self-efficacy instrument that is reflected in the finding of two factors. The categories are Self-efficacy in Test Taking, belief in ability in taking tests and Little Self-efficacy in Test Taking, or a lack of self-belief in taking tests. As well as the value of 0.32 for factor loadings, consistency of loadings, across the different conditions, onto one of these factors was also considered.

Costello and Osborne (2005) recommend dropping items that cross load if there are several adequate and strong loaders. Furthermore, they suggest that a factor with fewer than three items is weak. A loading of 0.5 is also seen as a strong factor loading. The following items have been assigned to the two factors according to value of 0.32 or more and for loading clearly and positively in four or more of the different conditions.

Factor 1(SETT) = Self-efficacy in Test-Taking: Q1 'If I find a test hard, I keep trying', Q2 'If I find a test hard, I still feel confident in my ability to pass', Q3 'If I did badly on a test, I still feel positive about other tests', Q10 'I can concentrate in tests even if I find them difficult', Q11 'I enjoy taking tests', Q12 'I can remain cheerful about life, even if a test is hard', Q13 'I can manage in tests, even if I feel tired or unwell', Q14 'I understand that even if I find a test difficult it is not the end of the world', Q15 'I don't worry about taking tests'.

Factor 2 (LSETT) = Little Self-efficacy Scale in Test taking: Q4 'If I find a test hard, I feel I will not pass', Q7 'If I find a test hard, I can't control negative thoughts about failing', Q8 'If I find a test hard, I start to doubt my ability to pass', Q9 'In tests, I panic if I see people finishing before me'.

Items Q5 'When I take tests, I feel anxious about being in a different room to my usual classroom' and Q6 'When I take tests, I feel anxious about the instructor' that showed skew and kurtosis in the descriptive statistics could be removed at this stage as they only loaded clearly and positively in three of the conditions. Item Q5 'When I take tests, I feel anxious about being in a different room to my usual classroom' was initially included to try and reflect the transactional nature (Zeidner & Matthews, 2005) of test anxiety; feelings that

may happen when taking a test outside of the usual classroom environment. Item Q6 'When I take tests, I feel anxious about the instructor' was included for the same reason.

A final test of Cronbach's *alpha* (see Table 7.20) with the new two subscales will inform the final decision whether these items, originally included to reflect the transactional 'state' rather than 'trait' (Spielberger & Vagg, 1995) test anxiety construct.

I am satisfied that I have fulfilled the criteria suggested by Brown (2015) for the exploratory factor process. I have chosen the correct estimator, determined the appropriate number of factors through scree plot and parallel analysis, used the correct rotation, interpreted the factors, and evaluated the quality of the solutions through eliminating poorly defined factors and poorly behaved items. I have not been able to test the factor structure on another sample.

Table 7.20

Cronbach's Alpha for the Self-Efficacy in Test Taking Instrument

Subscale and Items	Cronbach's *Alpha*
Self-efficacy in Test Taking: Q1, Q2, Q3, Q10, Q11, Q12, Q13, Q14, Q15	0.76
Little Self-efficacy in Test Taking: Q4, Q7, Q8, Q9	0.76
Little Self-efficacy Scale in Test Taking: Q4, Q7, Q8, Q9 and Q5 and Q6	0.76

Note. Proposed Subscales Across Both Intervention and Control Groups and at Pre-test and Post-test Timepoints. $N = 290$

The results show reliability for the Self-efficacy in Test-Taking subscale (SETT) and the Little Self-efficacy Scale in test-taking subscale. (LSETT)

Confirmatory factor analysis is different to exploratory factor analysis in that the model is driven by a theoretical approach rather than simply the data (exploratory factor analysis) and should be able to confirm the most parsimonious measurement model for the self-efficacy data.

Test anxiety theory, the idea of the transactional nature of test anxiety (Zeidner & Matthews, 2005) and Spielberger and Vagg's (1995) trait / state theory with Bandura's self-

efficacy theory (Bandura, 1997) are the theories that underpin the construction of the self-efficacy instrument. Exploratory factor analysis and tests for subscale reliability have not proved conclusive yet with regard to items Q5 and Q6. Confirmatory factor analysis will provide the final fit indices to determine the subscales to use in further analyses.

7.3.2.2 Confirmatory Factor Analysis of the Self-Efficacy Instrument. The same assumptions have been used as for the Children's Test Anxiety Scale Data detailed at section 7.2.2.1.

The Shapiro-Wilk test for Normality of Distribution (see Table H.5) confirmed that the data do not have a normal distribution as is typical of Likert type data. Therefore, in the final confirmatory factor analysis process, an "MLM" estimator with Satorra-Bentler method using Lavaan© in RStudio© was used. In addition, the sample size is not > 200 (Kline, 2015) and therefore does not meet the assumption (see note in Data Analytic Plan). For results (see Table 7.21).

Table 7.21

Fit Indices for the Proposed Model for the Self-efficacy in Test-Taking Scale N=290

Subscale	*Chi-Sq*	*df*	*p*	*RMSEA*	*pC*	*CFI*	*TLI*	*SRMR*
Self-efficacy in Test-Taking: Q1, Q2, Q3, Q10, Q11, Q12, Q13, Q14, Q15	56.39	27	.01	0.06	.20	0.93	0.91	0.05
Little Self-efficacy Scale in Test-Taking: Q4, Q7, Q8, Q9	0.64	2	.73	0.00	.90	1.00	1.02	0.01
Little Self-efficacy Scale in Test-Taking: Q4, Q7, Q8, Q9 and with Q5 and Q6	50.18	9	.00	0.13	.00	0.89	0.81	0.07

Note. X^2 = *Chi-square; df = degrees of freedom; RMSEA = Root Mean Square Error of Approximation; pC = pCLOSE* = value of *RMSEA; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; SRMR = Standardized Root Mean Square Residual*

The results show that the Self-efficacy in Test Taking subscale (SETT) is a good fit to the data. Goodness of fit indices *CFI* and *TLI* are 0.93 and 0.91 respectively. The badness of

fit *RMSEA* is at an excellent level of 0.06. (Hu & Bentler,1999). The *SRMR,* should ideally be near to 0.00 as possible is good at a value of 0.05.

The Little Self-efficacy in Test Taking subscale (LSETT) with items Q5 and Q6 removed is an excellent fit. Goodness of fit indices *CFI* and *TLI* are 1.00 and 1.02 respectively. Badness of fit index *RMSEA* is 0.00 indicating a perfect fit and *SRMR* is also 0.01 indicating an almost perfect fit. There are only four items on this subscale, but the guidelines suggest no fewer than three (Brown, 2015).

However, the Little Self-efficacy in Test Taking subscale when tested with items Q5 and Q6, identified for their skew and kurtosis in the descriptive statistics are included, there is a much poorer fit (see Table, 7.21). Goodness of fit indices *CFI* and *TLI* are 0.89 and 0.81 respectively and therefore do not meet the established criteria of as near to 0.95 as possible. The badness of fit index *RMSEA* is very high at 0.13; the ideal being around 0.06. Including Q5 and Q6 into the LSETT subscale results in a bad fit and therefore these two items will be dropped from the Self-efficacy in Test Taking instrument resulting in a thirteen-item instrument consisting of two subscales (see Figure 7.2, Figure 7.3)

Self-efficacy in Test-Taking Instrument subscales:

Self-efficacy in Test Taking (SETT) = Q1, Q2, Q3, Q10, Q11, Q12, Q13, Q14, Q15

Little Self-efficacy Scale in test taking (LSETT) = Q4, Q7, Q8, Q9

Figure 7.2

Self-efficacy in Test Taking Scale

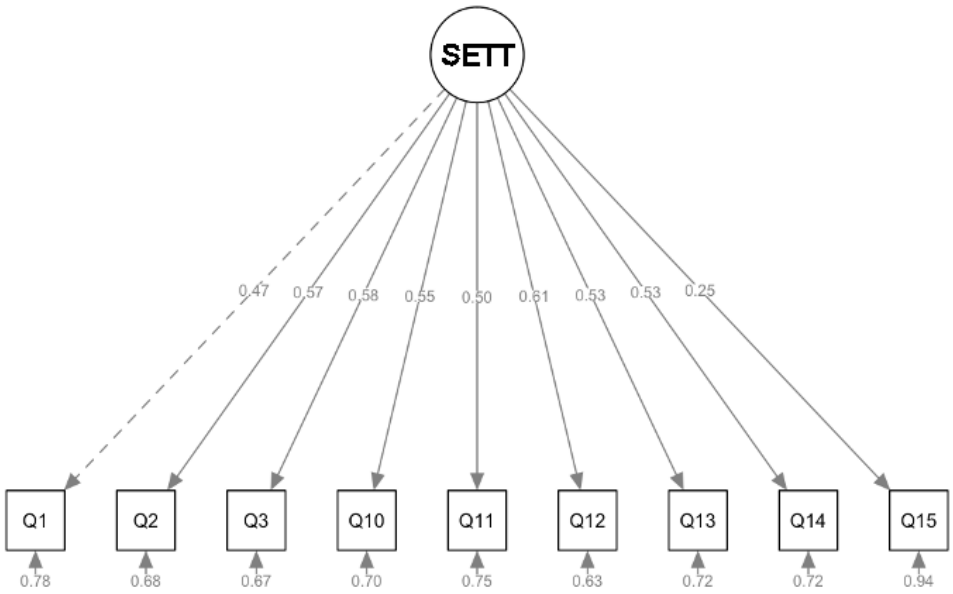
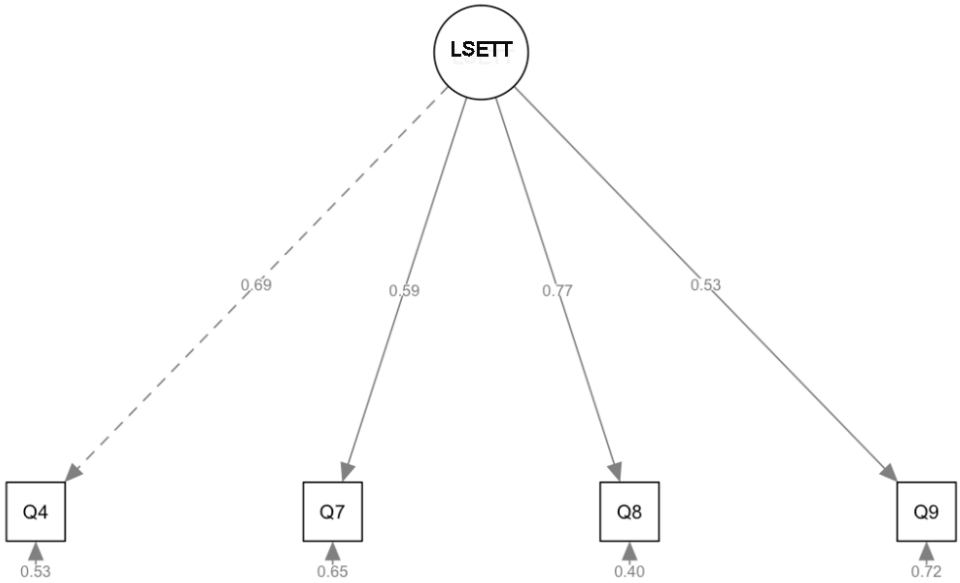


Figure 7.3

Little Self-efficacy in Test Taking Scale



To consider taking this model forward for further analyses to answer the research questions, it is useful to use multigroup confirmatory analysis to complete measurement invariance of the model across timepoints (Pre-test and post-test), groups (intervention and control) and gender (boys and girls) to check that the construct does not behave differently in different conditions. This process is the same as was used for the Adapted Thoughts' subscale model and will ensure that the most parsimonious model is used for both the Children's Test Anxiety Scale data and the Self-efficacy in Test Taking data in this study.

7.3.3 Measurement Invariance for the Self-efficacy Test Taking Instrument (SETTI)

The two subscales of the Self-efficacy in Test Taking Instrument, identified through the exploratory factor analysis and confirmatory factor analysis processes, were tested for measurement invariance across timepoint, intervention and control groups and across gender groups. Where measurement invariance holds from configural to scalar invariance, further analysis of the latent means is possible. (Brown, 2015)

Measurement invariance held for the Self-efficacy in Test Taking (SETT) and for the Little Self-efficacy Scale in test taking (LSETT) for the intervention and control groups and across the timepoint condition pre-test to post-test at the configural to constrained model comparison stage.

For measurement invariance across the intervention and control groups, partial measurement invariance is held for the Self-efficacy in Test Taking subscale (SETT) at the metric stage by controlling for Question 11 'I enjoy taking tests', and measurement invariance holds at the partial metric to scalar stage. For the Little Self-efficacy Scale in Test Taking, measurement invariance holds at the constrained to metric stage.

For measurement invariance across the gender groups, boys and girls, partial measurement invariance holds for the Self-efficacy in Test Taking (SETT) subscale at the metric stage, controlling for Question 10 'I can concentrate in tests even if I find them difficult' and measurement invariance holds at the partial metric to scalar stage. For the Little Self-efficacy Scale in test taking (LSETT) subscale, measurement invariance holds at the configural and constrained comparison of model stage. All results and *chi square* differences between stages (see Tables H.11 – H.23)

Measurement invariance as a precursor to latent means analysis is essential in establishing whether there is non-invariance. Measurement invariance or partial measurement invariance holds across pre-test and post-test timepoint groups, intervention, and control

groups and across boy and girl gender groups. This result is important when establishing a new scale (Brown, 2015; von der Embse et al., 2021)

Brown (2015) suggests that comparison of latent means in confirmatory factor analysis is comparable to, but stronger than, ANOVA. This comparison of means is being conducted in the context of the measurement model. Latent means and the more traditional ANOVA are used for this question in this study.

7.4 Intervention Data

The next section adds to the Data Analytic Plan by outlining how I meet the assumptions for the proposed analyses and the sequential decisions that inform the final analyses in the results chapter.

7.4.1 Assumptions

Mixed, 2 x 2 factorial ANOVAs are used for Research Question 1, Research Question 2a and Research Question 3. Like all other parametric tests, a mixed 2 x 2, factorial ANOVA makes a series of assumptions that should be met before the analyses. The *within-subjects* factor should contain at least two related (repeated measures) categorical groups (levels) and in this case this factor is 'Timepoint' Pre-test and post-test. The *between-subjects* factor should have at least two categorical independent groups (levels) and in this case this factor is the intervention and control groups. The dependent variable should be continuous and approximately normally distributed for all combinations of factors and in this case, this measure is the response values of the participants' sum scores on the Likert type scales. There should be homogeneity of variance for each of the groups and, if more than two levels, sphericity between the related groups. Homogeneity of variance was tested using Levene's test and found to be non-significant for the Children's Test Anxiety Scale data and for the Self-efficacy data for all proposed analyses. Therefore, the null hypothesis that the variance is equal across the groups could be accepted. I did not need to use the sphericity test as there were not more than two levels for any of the questions (see Appendices G and H).

Finally, there should be no significant outliers, and these were removed through the exploratory factor analysis and confirmatory factor analysis processes. (Field, 2012)

The sample sizes meet the power requirements for Research Questions 1, 2a, 2b, 3, 4 and 5 (see 5.3.4.1).

7.4.1.1 Assumptions for ANCOVA. These are the same as for the ANOVAs but with the addition of independence of the covariate and homogeneity of regression slopes (Field, 2012). For part of Research Question 2a that was answered with the Little Self-efficacy Subscale, these assumptions were met. (see Figure 8.8)

7.4.1.2 Assumptions for Multi-level Linear Modelling. These are the same as for most linear models, linearity, additivity, homogeneity of variance. In addition, there should be normal distribution of the random intercepts. (Field, 2012). Sample size should be as large as possible (Field, 2012) and level two units should be approximately fifteen. However, Gelman and Hill (2006) have suggested as few as two level two units depending upon software used. There are 11 level two units for Research Question 1. Sample size will also depend on the model being fitted, in this case, it is the random intercepts model. (e.g., Browne, n.d; Sample Sizes for Multilevel Models). Twisk (2006) recommends using sample size calculations with caution. Kreft and de Leeuw (1998) suggest that there are so many factors involved in multilevel linear analysis that it is impossible to produce meaningful rules about sample size and power.

For Research Question 1, where the homogeneity of slopes assumption was not met and therefore an ANCOVA could not be used, multilevel linear modelling examined clusters of data through a comparison of different models that allowed the intercepts to vary and used two level 1 fixed effect predictors of group and the pre-test score.

Multilevel linear modelling allows a researcher to look at the data hierarchically. In this case, I looked at the data for schools as Level 2 in the hierarchy and the students as Level 1. I looked to see if any of the explanatory variables, such as intervention or control group and the pre-test scores affected different schools in different ways. The multilevel linear model is especially important in educational research as it allows for the data to be viewed at the school level. (e.g., Rasbach, n.d; What are multilevel models and why should I use them)

Multilevel linear modelling partitions the residual variance into a between school component and the within school variance that looks at the variance for each child. These unobserved school characteristics may affect child outcomes. The random intercept model allows each school group to have a different slope (Pillinger, n.d; Random intercept models)

Multilevel linear modelling allows for heterogeneity amongst the groups through the variance partition coefficient (Brunton-Smith, 2019). The variance components model is used to provide the initial estimation of the contribution of variance that was from the different

groups and to determine if it is worthwhile running a multilevel linear model. Intraclass correlation values above 5% show that there is clustering and that it is worth looking at the data hierarchically (Heck et al., 2014)

Multilevel linear modelling allows a researcher to ask 'what if' questions. What happens if we move the intercepts of the responses from the data and allow the slopes to vary? Removing the constrictions of traditional data interpretation and allowing for group (school or class) hierarchical interpretation can help a researcher understand the data better and as a result of this process understand, where any problems in the design of data collection may lie and gather a true picture of what is happening in the data.

Field (2012) and Heck et al., (2014) recommend centering the level 1 predictors to stabilise the model. This centering was done for the level 1 explanatory variable of pre-test scores for both the intervention and control group using a grand centred mean rather than a group centred mean as I was interested to see what happened at level 2 with the schools (Field, 2012).

The multilevel linear model was calculated using SPSS © software and to ensure the random effects assumption held, maximum likelihood was used as an estimator. In addition, the variance components option was used as the coefficient matrix and *Wald z* is reported (Antonakis et al., 2019). For results, (see Appendix G, Tables G.37 – G.49)

For research question 2a where there is also non-homogeneity of slopes, I used the variance components calculation to see whether I should run the multilevel model but there was not enough variance in the data to make this analysis worthwhile (Field, 2012).

7.4.1.3 Post-hoc Tests. Post-hoc tests use pairwise comparisons to control for familywise error by correcting the level of significance for each test so that the overall Type 1 error remains at *alpha* = 0.5. The *alpha* value is then divided by the number of comparisons (Field, 2012).

For part of Research Question 2a (Little Self-efficacy Scale in Test Taking subscale) and for Research Question 3, where the ANOVAs showed a significant interaction, *Tukey's Honestly Significant Different* post-hoc test was used as homogeneity of variance was established through Levene's test for both analyses.

7.4.1.5 Assumptions for the Independent *T test* (RQ 5). The usual tests for sources of bias apply, thus, additivity, linearity, homogeneity of variance and normality: parameter estimates, confidence intervals, errors, null hypothesis significance testing, (Field, 2012).

Levene's test showed homogeneity of variance for the online versus paper samples used to answer Research Question 5.

7.4.1.6 Assumptions for Latent Means Analyses (RQ1, RQ2a). Traditionally, researchers who want to compare group means use an ANOVA with an observed sum score (Finch & French, 2015). However, there is an assumption that the item responses are made without error which is often not the case (Brown, 2015). Relying on latent means analyses to compare group means allows one to correctly account for measurement error, if full invariance holds across the groups being compared. Finch and French (2015) consider the assumption for full invariance to be conservative and suggest that latent means analysis is possible where partial invariance holds across the groups being considered.

Given that full invariance holds for the Adapted Thoughts' subscale across groups, gender and timepoint conditions and full invariance holds across timepoint conditions for both subscales of the Self-efficacy in Test Taking instrument, and partial invariance for both subscales across gender and group, latent means analyses has been calculated alongside the more traditional ANOVAs. The effect size calculation comes from Hancock (2001). (see Appendix G, Tables G.33 – G.36; Appendix H, Tables, H.28- H. 31; H.40 – H.43)

7.4.1.7 Assumptions for MIMIC (Multiple Indicator Multiple Causes) models. Q2b. MIMIC models can provide rigorous results (Chang et al., 2020). Diemer (2019) queries why they are not used more often. The assumptions for running a MIMIC model come from the goodness of fit statistics from confirmatory factor analysis for the latent factor and an existing dichotomous observed covariate or a continuous observed covariate. The Adapted Thoughts' subscale is the latent factor, and its goodness of fit statistics are reported in Adaptation 3 (see Table 7.15). The individual sum scores from the two subscales, Self-efficacy in Test Taking and Little Self-efficacy subscales are the continuous, observed covariates used.

8. Results

In this chapter, the results are presented simply and often graphically. The results and their links to ideas in the literature review and possible implications for research are discussed in Chapters 9 and 10.

Tables for all results are in appendices G and H. The assumptions for factor analysis and further analyses, as detailed in the Data Analytic Plan in Chapter 7 are not reported again here. Levene's test for homogeneity of variance was non-significant for all of the research questions where it was applicable. (see Tables G.29, G.50, G.54, H.24, H.32, H.44, H.49, H.53). Furthermore, for Research Questions 1and part of 2a, where the assumption of homogeneity of slopes for ANCOVA was not met, multilevel linear analyses were tried.

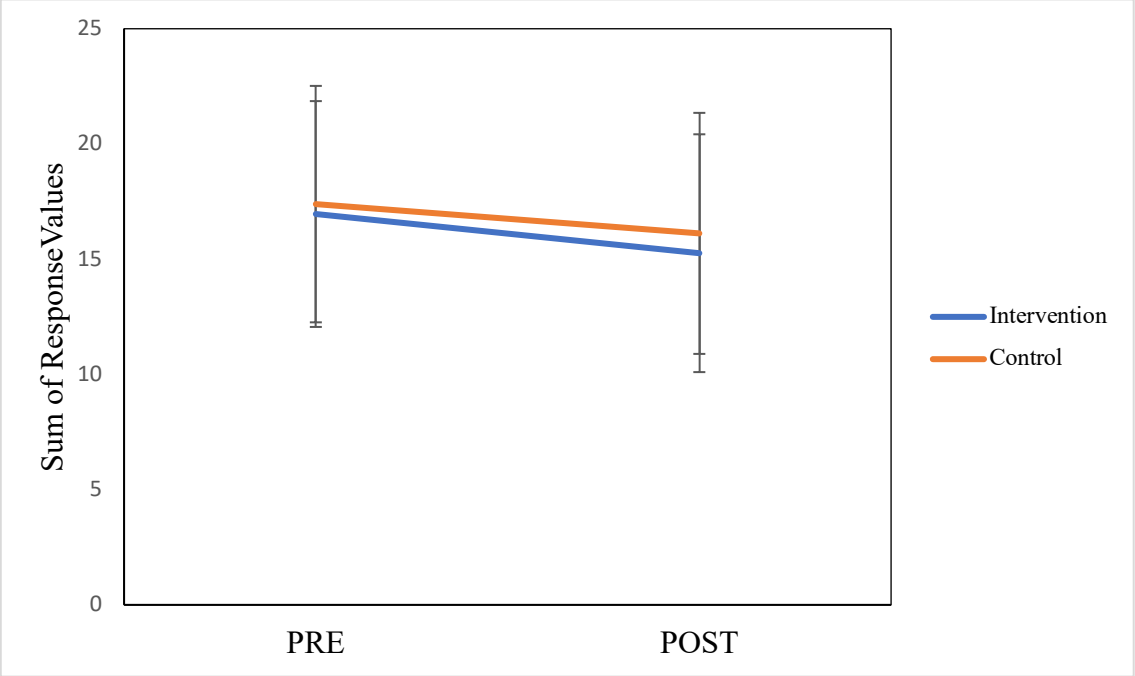
When reporting the results of the ANOVAs, *partial eta* effect size is taken from Cohen, (1988) with values of: small effect = .01, medium effect = .09 and a large effect = 0.25. Graphs use standard error bars. For the latent means results that are reported alongside the ANOVAs, for reasons of interest only, effect sizes were calculated from Hancock (2001).

8.1 The Children's Test Anxiety Scale Results

For Research Question 1, a 2 x 2 mixed factorial ANOVA was run with the independent variables of timepoint and group. The dependent variable was the participants' sum score from the *Adapted Thoughts'* subscale. There was a significant main effect for the independent variable Timepoint, F (1,168) = 23.55, $p < .001$, $\eta^2_p = 0.12$ with pre-test scores (M = 16.96, SD = 4.90) being higher than post-test scores (M =15.26, SD = 5.16) for the intervention group and for the control group pre -test scores are higher (M = 17.39, SD = 5.13) than post-test scores (M = 16.12, SD = 5.23). The effect was not significant for the independent variable Group, F (1,168) = 0.79, p = .37, $\eta^2_p = 0.005$. There was no interaction between Timepoint and Group, F (1,168) = 0.50, p = .48, $\eta^2_p = 0.003$. (see Figure 8.1) Mean values and standard deviations are reported (see Appendix G, Tables G.29 - G.32)

Figure 8.1

Means of Pre-test and Post-test Intervention and Control Groups for the Children's Test Anxiety Scale Data Measured Using the Adapted Thoughts' Subscale



8.1.1 Latent Means Analysis

For the intervention groups, the pre-test latent means was compared to the post-test latent means and the result was significant, $p = .01$. The effect size of the difference was calculated. The effect size, $ES = 0.37$ is comparable to Cohen's effect sizes (Hancock, 2001) and is therefore approaching a middle effect size. (Cohen, 1988)

For the control groups at pre-test and post-test timepoint, the result was non-significant, $p = .09$. Although, the ANOVA between the pre-test and post-test for the control groups is a non-significant result at $p = .09$, the calculation reveals that there is a small effect size of $ES = 0.25$. (see Appendix G, Tables G.33 – G.36)

8.1.2 Multilevel Linear Analysis for the Children's Test Anxiety Scale

The assumption for the homogeneity of slopes for an ANCOVA was not met and thus led to exploring the data hierarchically to see if there are groups (schools) within the data that have clusters. The patterns in the scatterplots (see Figures 8.2 and 8.3) suggest that there are clusters of pre/post-test scores by school.

Figure 8.2

Pre-test as a Covariate to the Post-test Scores for the Intervention and Control Groups Showing Non-homogeneity of Slopes and Clustering

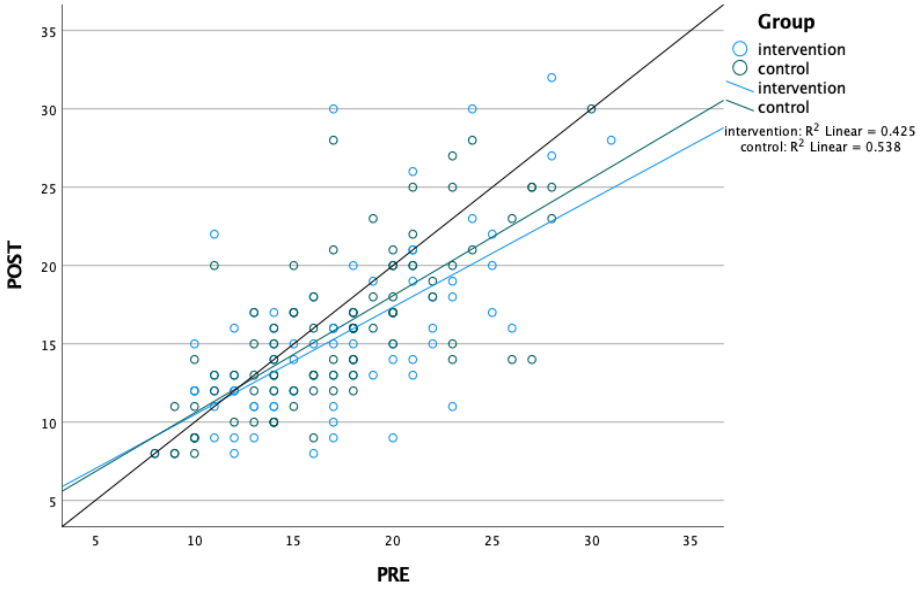
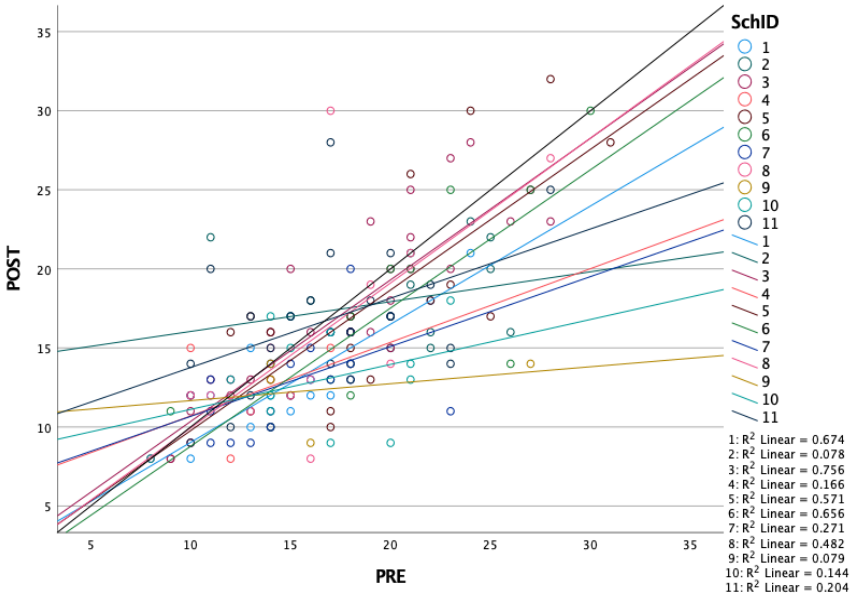


Figure 8.3

Pre-test as a Covariate to the Post-test Scores for the Different Schools



The variance components model showed an Intraclass Correlation of 11%. This result suggests that approximately one school in the data is contributing to the clusters but for 89% of the data, the clusters are at participant level. When the random intercepts model with group and pre-test centred means as level 1 fixed effects predictors was fitted, there was a very significant change, $\Delta X^2 = < .01$ in the fit of the model. At this stage, this result means that by allowing the pre-test centred mean score's intercepts to vary, the model is a better fit to the data.

When the pre-test scores are allowed to randomly intercept the slopes, the variance between the schools, measured by the Intraclass Correlation was reduced from 11% to 5%. (see Appendix G, Tables G.37 -G.49)

8.2 Self-efficacy in Test Taking Results

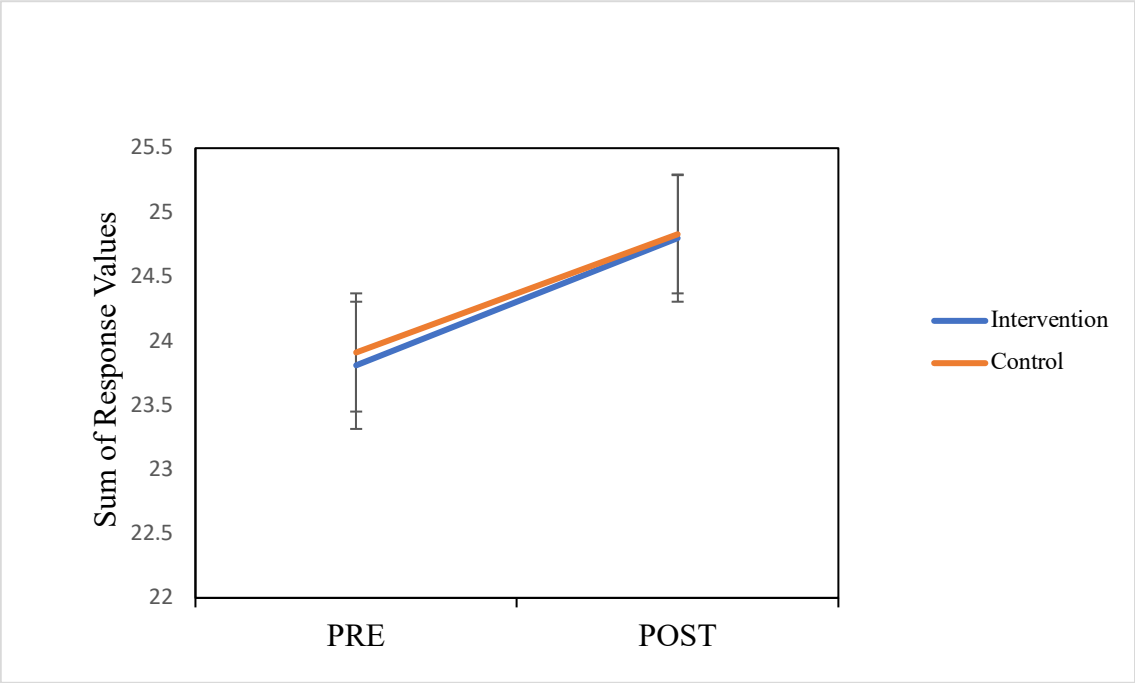
Research Question 2a is answered using both the Self-efficacy in Test Taking Scale and the Little Self-efficacy Scale in Test Taking scale. If self-efficacy in test taking increases, one would expect to see an increase in means pre-test to post test in the Self-efficacy subscale and a decrease in means pre-test to post-test in the Little Self-efficacy Scale subscale.

For Research Question 2a, a mixed 2 x 2 Factorial ANOVA on the Self-efficacy in Test Taking Instrument Subscale was run with the independent variables of Timepoint and Group. The dependent variable was the sum score on the Self-efficacy in Test Taking subscale.

There was a significant main effect for Timepoint, F (1,143) = 5.25, p = .02, η^2_p = 0.04 with pre-test scores (M = 23.81, SD = 4.32) being lower than the post-test scores (M = 24.78, SD = 5.00) for the intervention group and for the control group, pre-test scores (M = 24.31, SD = 5.06) are lower than post-test scores (M = 24.78, SD = 4.99). There was no significant effect for the independent variable Group, F (1,116) = 0.400, p = .53, η^2_p = 0.003 and no interaction between the Timepoint and Group, F (1,116) = 1.56, p = .21, η^2_p = 0.013. (see Figure 8.4; Appendix H, Table H.24 – H.27)

Figure 8.4

Means of Pre-test and Post-test Intervention and Control Groups for the Self-efficacy in Test Taking Data Measured Using the Self-efficacy in Test Taking Subscale



8.2.1 Latent Means Analyses

A comparison of the latent means was conducted on the intervention group and the control group from pre-test to post-test and their effect sizes calculated according to Hancock (2001) reported here.

For the Self-efficacy in Test Taking subscale across the intervention groups, $p = .16$, a non-significant effect. The effect size calculation was $ES = .24$. For the Self-efficacy in Test Taking across the control group, $p = .20$ and was non-significant. The effect size calculation was $ES = .20$. (see Appendix H, Tables H.28-H.31)

8.2.2 Mixed 2 x 2 Factorial ANOVA on the Little Self-efficacy Scale in Test Taking Subscale

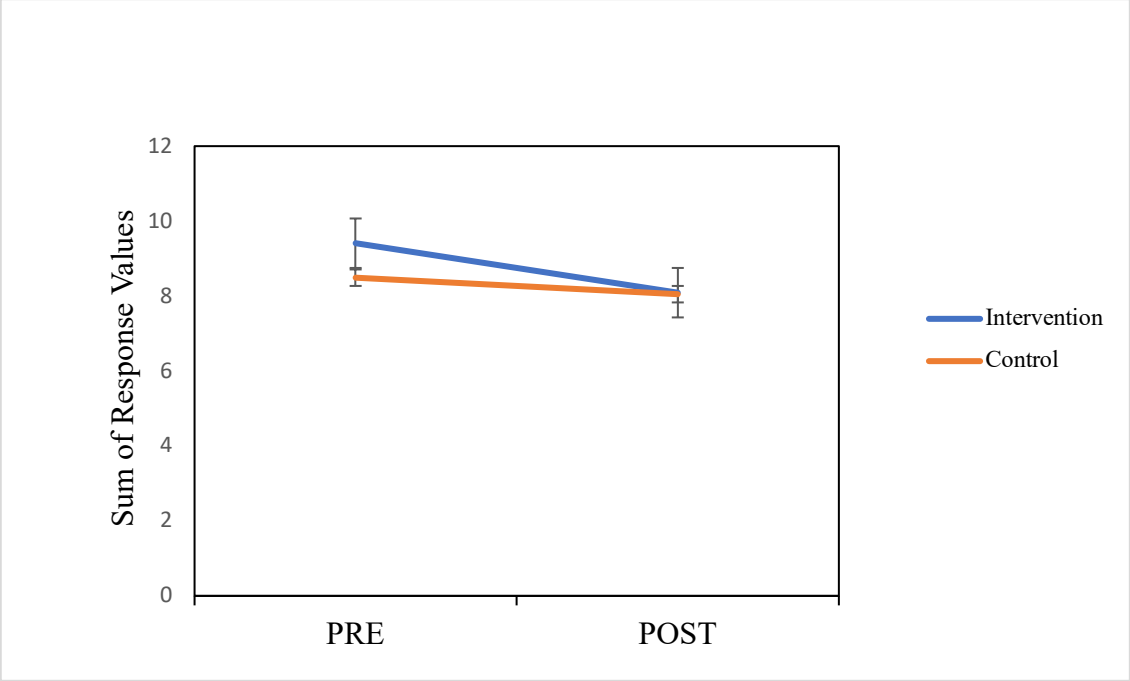
The second and smaller subscale of the Self-efficacy in Test Taking Instrument measures negative feelings of efficacy about taking tests.

There was an interaction between the Groups and the Timepoint, $F(1, 143) = 0.12$, $p = .04$, $\eta^2_p = 0.03$. with pre-test scores for the intervention group ($M = 9.41$, $SD = 4.32$) being

higher than the post-test scores (M = 8.09, SD = 4.99) and the control group pre-test scores (M = 8.49, SD = 5.01) being higher than the post-test scores (M = 8.05, SD = 5.17). Post-hoc *Tukey's Honestly Significant Difference (HSD)* revealed that there was a significant increase in feelings of self-efficacy in test taking in the intervention group, $p < .001$. (see Figure 8.5; Appendix H, Table H.32 - H.36)

Figure 8.5

Means of Pre-test and Post-test Intervention and Control Groups for the Little Self-efficacy Scale in Test Taking Data Measured Using the Little Self-efficacy in Test Taking Subscale



8.2.3 Latent Means Analyses

A comparison of the latent means was conducted on the intervention group and the control group from pre-test to post-test and their effect sizes calculated according to Hancock (2001). For the Little Self-efficacy in Test Taking subscale across the intervention groups, $p < .001$. There is a significant difference from pre-test to post-test in the intervention group. The effect size calculation was ES = .47. For the Little Self-efficacy in Test Taking subscale across the control group, p = .23 and was non-significant. The effect size was ES = -.21. (see Appendix H, Tables H.40 – H.43)

8.2.4 Multilevel Linear Analysis for the Self-efficacy in Test Taking Scale

The assumption for the homogeneity of slopes, shown in figures 8.5 and 8.6 to conduct an ANCOVA was not met and thus led to considering multilevel linear analysis as per section 8.2.1. However, there was not enough variance in the variance components model to calculate an initial Intraclass Correlation and to continue with a multilevel model would be mental flagellation (Field, 2012).

Figure 8.6

Pre-test as a Covariate to the Post-test Scores for the Intervention and Control Groups Showing Non-homogeneity of Slopes

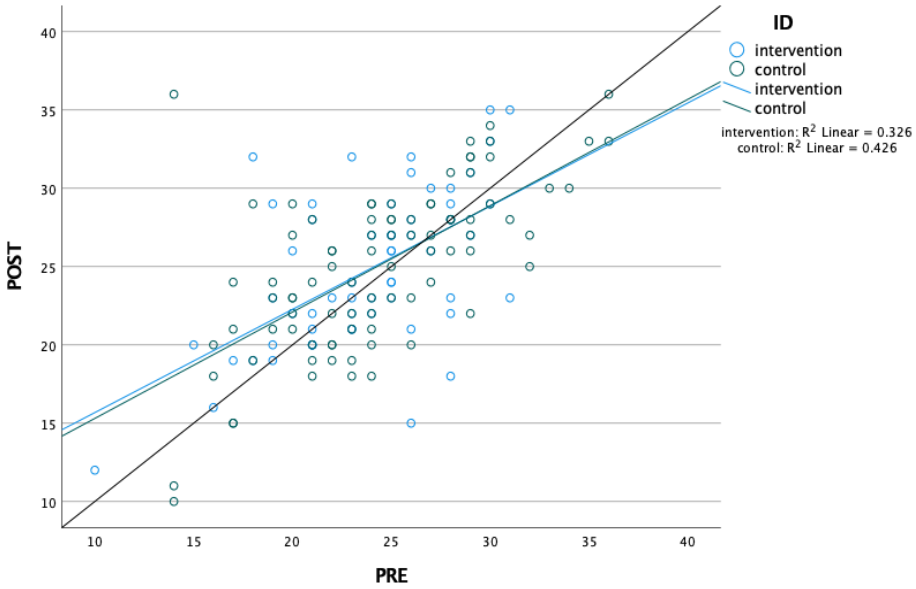
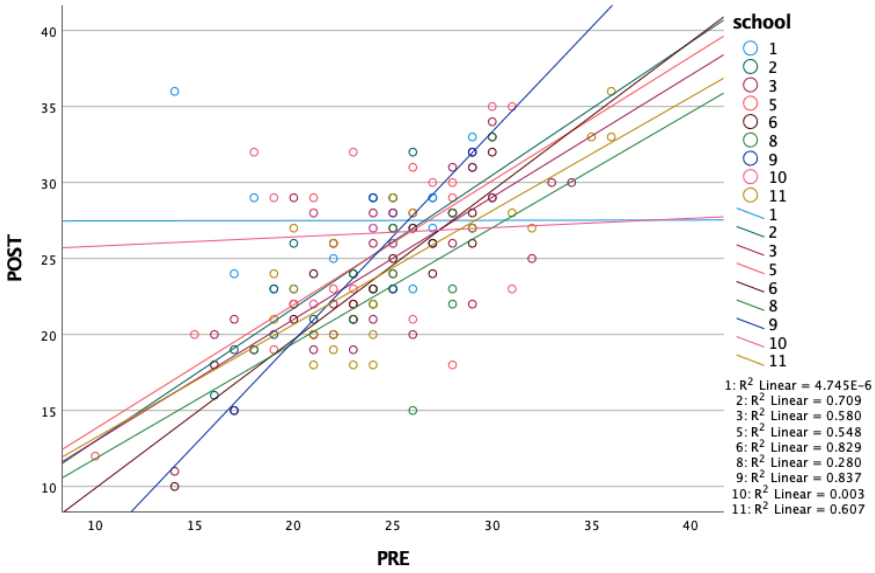


Figure 8.7

Pre-test as a Covariate to the Post-test Scores for the Different Schools



8.2.5 ANCOVA for the Little Self-efficacy in Test Taking Scale

The assumption for homogeneity of slopes is met (see Figure 8.8, Figure 8.9). A One-way ANCOVA was conducted to determine a statistically significant difference between the intervention and the control group on the post-test score controlling for the pre-test score. There is a not a significant effect of group on post-test score when controlling for the pre-test score, F (1,142) = 1.95, p = 0.16, η^2_p = 0.014. (see Appendix H, Tables H.37 – H.39)

Figure 8.8

Little Self-efficacy in Test Taking Subscale Groups

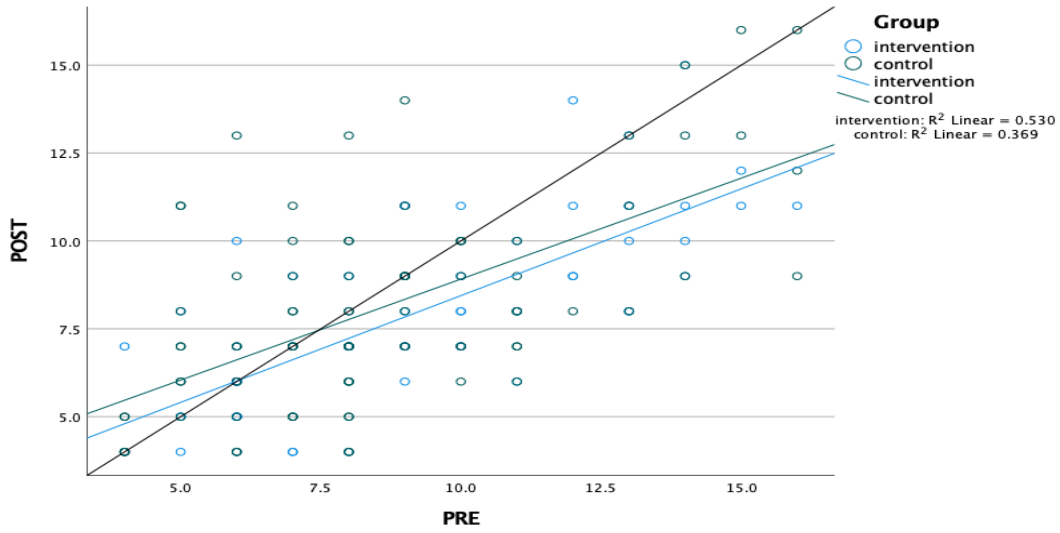
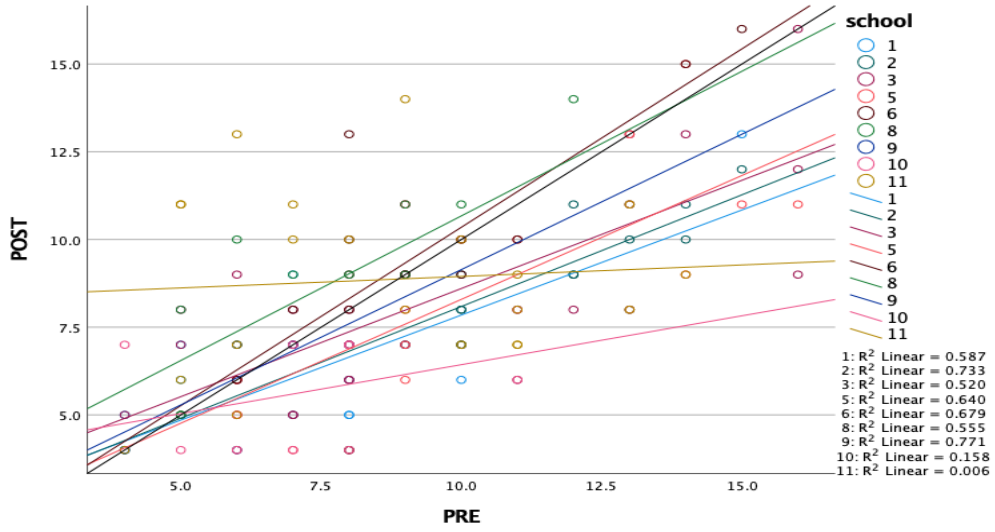


Figure 8.9

Little Self-efficacy in Test Taking Schools



8.3 Research Question 2b

8.3.1 Multiple Indicators Multiple Causes – MIMIC Models

Covariate significance values and estimates are reported here (see Table 8.1). *Z* scores and standard error are reported at Appendix G, Table G.57 and Figures G.17 -G.28. The estimate represents the slope in a regression as the covariate is a continuous variable (Finch & French, 2015). The model is represented (see Figure 8.10)

Figure 8.10

The MIMIC (Multiple Indicators Multiple Causes) Model (Diagram from Finch and French, 2015) for the Adapted Thoughts Latent Variable with Self-efficacy in Test Taking Subscales as Exogenous Variables (Covariate)

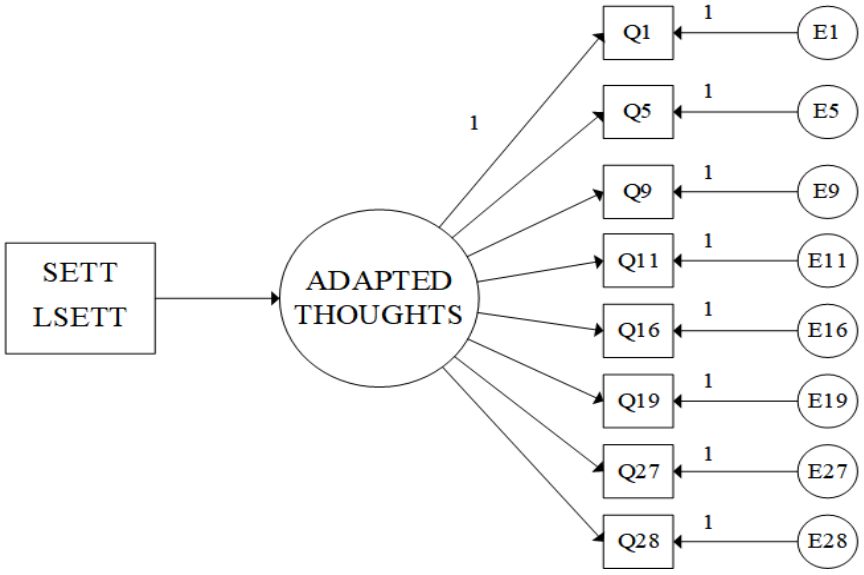


Table 8.1

Results of Multiple Indicators Multiple Causes(MIMIC) Models for the Adapted Thoughts' Subscale

Timepoint	*Covariate*	*Condition*	*Scale*	*p-value*	*estimate*
Pre-test	Pre-test Score	Intervention	Self-efficacy	.01	-0.04
			Little Self-efficacy	.006	0.07
		Control	Self-efficacy	.68	-0.001
			Little Self-efficacy	.69	0.002
Post-test	Pre-test Score	Intervention	Self-efficacy	.04	-0.04
			Little Self-efficacy	.001	0.12
		Control	Self-efficacy	.29	-0.01
			Little Self-efficacy	.24	-0.01
Post-test	Post-test Score	Intervention	Self-efficacy	.03	-0.04
			Little Self-efficacy	.001	0.10
		Control	Self-efficacy	.56	0.008
			Little Self-efficacy	.10	0.01

If the estimate is statistically significant, one can conclude that the covariate (Self-efficacy in Test Taking or Little Self-efficacy in Test Taking) is related to the factor (Finch & French, 2015), the Adapted Thoughts' subscale that measures the worry element of test anxiety. For example, in the pre-test intervention self-efficacy, the estimate is -0.04 (which in this case shows the negative direction of the slope) and shows that students who are high in

self-efficacy are related to low worry and students who are low in self-efficacy are related to high worry.

8.4. Students Who Are High in Test Anxiety

In this study the mean score for the pre-test intervention group from the Children's Test Anxiety Scale was $M = 62.85$ and the standard deviation, $SD = 15.15$. Therefore, students, who had a sum score $M = 78.00$ or over (*Mean* plus one *standard deviation*) were students who were high in test anxiety. Those who had a sum score, $M = 47.7$ (*Mean* less one *standard deviation*) or less were students with low test anxiety.

For the control pre-test, group the mean score was $M = 62.26$ and the $SD = 15.18$. Therefore, a student with a sum score, $M = 77.44$ (*Mean* plus one *standard deviation*) are students with high test anxiety and a student with a score, $M = 47.08$ (*Mean* less one *standard deviation*) are students who have low test anxiety.

8.4.1 Research Question 3

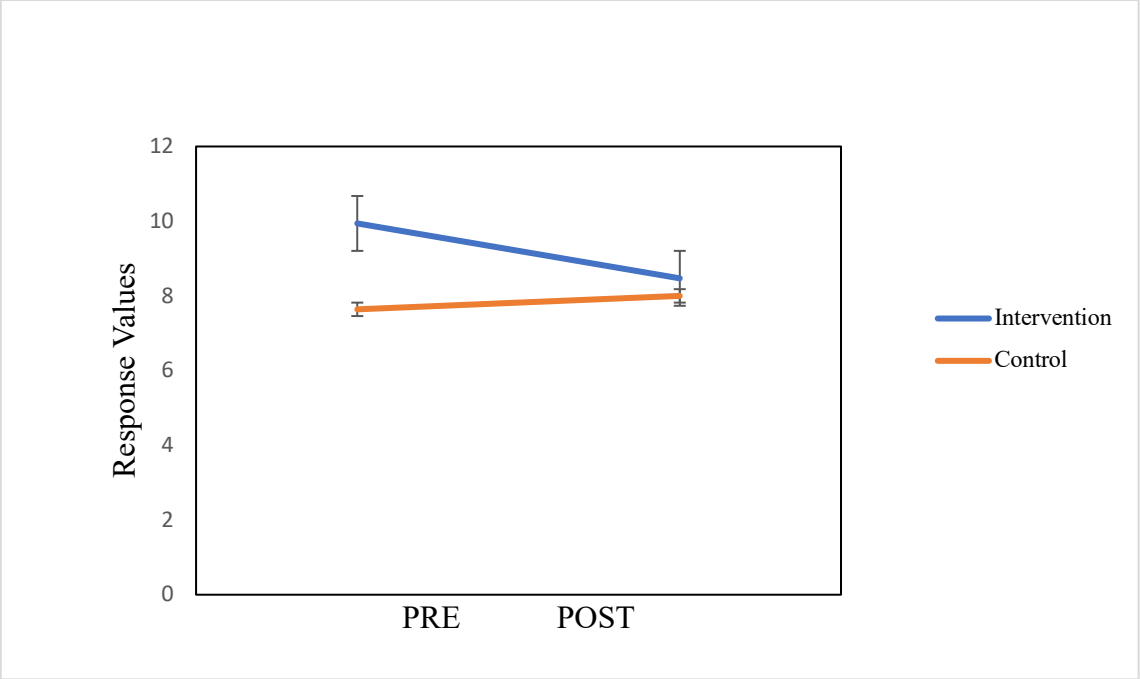
A mixed 2 x 2 ANOVA was run for students who are high in test anxiety using the Little Self-efficacy Scale in Test Taking Subscale, the Self-efficacy in Test Taking subscale and finally the Adapted Thoughts' subscale.

For the Little Self-efficacy in Test Taking subscale, there was a significant interaction between group and timepoint, $F(1,29) = 5.57$, $p = .03$, $\eta^2_p = 0.07$. The η^2_p effect size is between a medium and a large effect. Post-hoc *Tukey's Honestly Significant Difference (HSD)* revealed that there was a significant increase in feelings of self-efficacy in test taking in the intervention group, $p = .04$.

For the intervention group that consisted of students, who are high in test anxiety, means reduced from pre-test ($M = 9.94$, $SD = 3.56$) to post-test ($M = 8.47$, $SD = 3.04$) to show an improvement in feelings of self-efficacy after the intervention. The mean for the students, who are high in test anxiety, for the control group, increased from pre-test ($M = 7.64$, $SD = 2.53$) to post-test ($M = 8.00$, $SD = 2.22$) showing a reduction of feelings in self-efficacy after the test. (see Figure 8.11; Appendix H, Tables H.44 – H.48)

Figure 8.11

Means of Pre-test and Post-test Students who are High in Test Anxiety for the Intervention and Control Groups Data Measured Using the Little Self-efficacy Scale in Test Taking Subscale



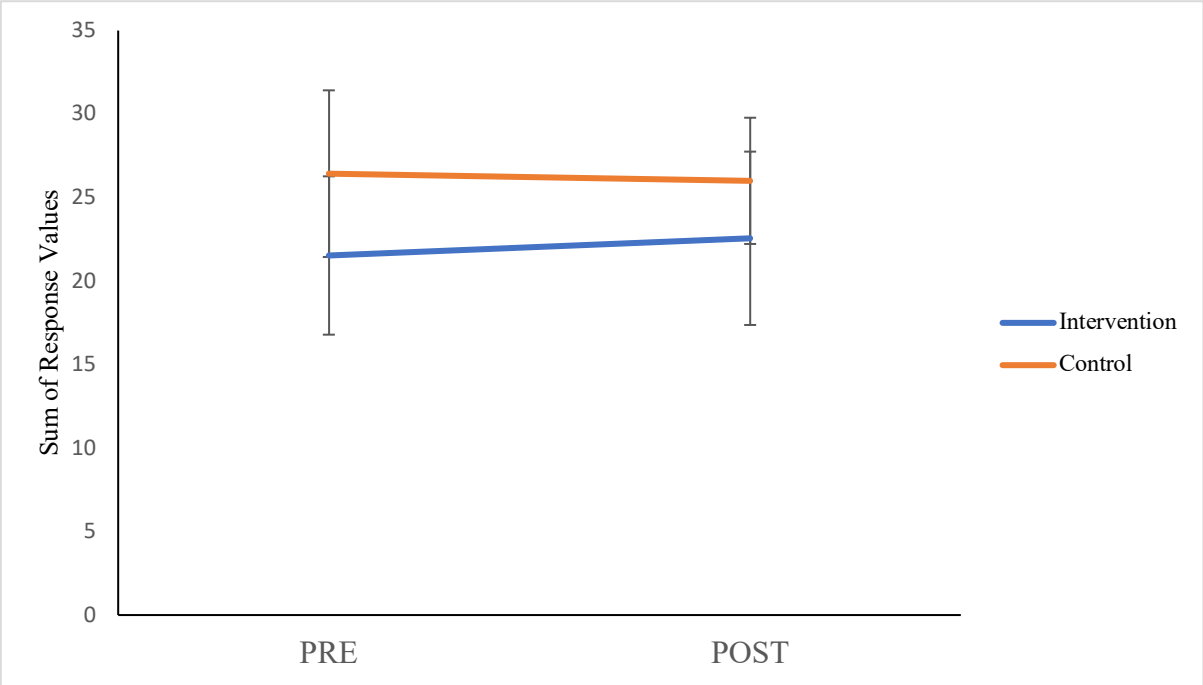
A further mixed 2 x 2 factorial ANOVA was run with the repeated measure variable at pre-test and post-test, Timepoint and the between Group variable as the intervention group or control group. The dependent variable was taken from the total of the Self-efficacy in Test Taking subscale.

There was a significant main effect for Group, F (1,29) = 7.36, p = .01, η^2_p = 0.20 with the intervention group pre-test means increased from pre-test (M = 21.53, SD = 4.74) to post-test (M = 22.56, SD = 5.19) showing an increase in feelings of self-efficacy in test taking and the control group means decreased from pre-test (M = 26.43, SD = 4.99) to post-test (M = 26.00, SD = 3.78) showing a reduction in feelings of self-efficacy. (see Figure 8.12; Appendix H, Tables H.49 -H.52)

There was no significant main effect for Timepoint, F (1, 29) = 0.15, p = .70, η^2_p = 0.005. The interaction between group and timepoint was not significant, F (1,29) = 0.94, p = .34, η^2_p = 0.03.

Figure 8.12

Means of Pre-test and Post-test Students who are High in Test Anxiety for the Intervention Groups Data Measured Using the Self-efficacy in Test Taking Subscale



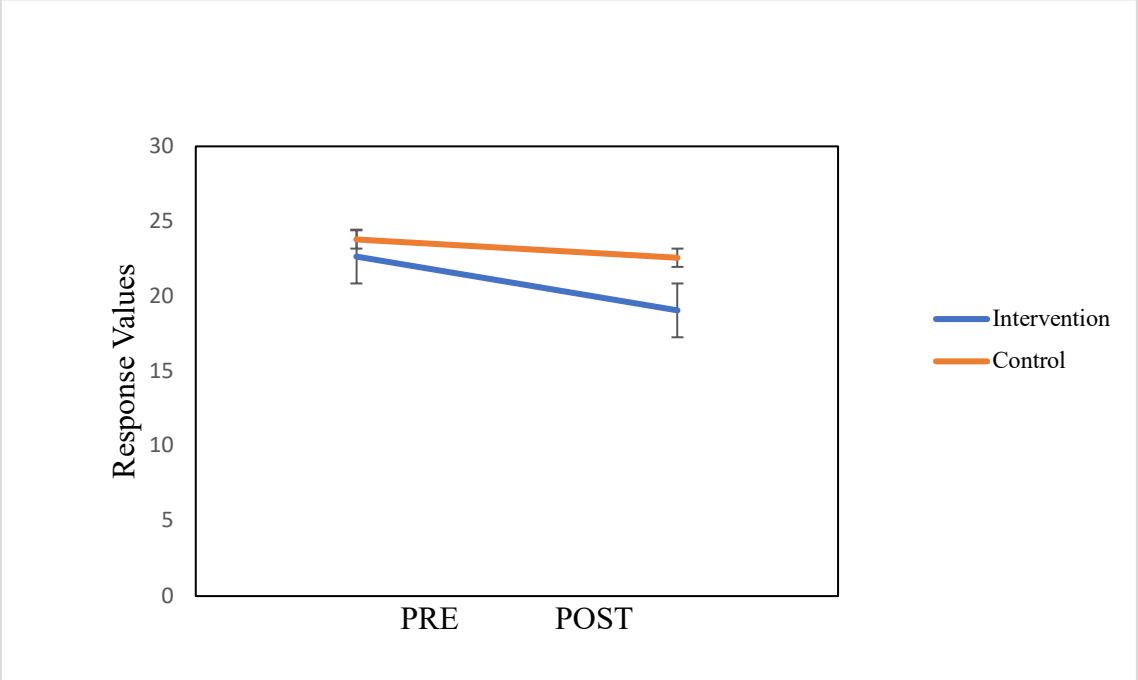
A further mixed 2 x 2 factorial ANOVA was run with the repeated measure variable at pre-test from the Adapted Thoughts' subscale and post-test from the Adapted Thoughts' subscale, Timepoint and the between Group variable as the intervention group or control group. The dependent variable was taken from the total of the Adapted Thoughts' subscale.

There was a significant main effect for Timepoint, F (1,29) = 7.75, p = .009, η^2_p = 0.21 with the intervention group pre-test means decreased from pre-test (M = 22.65, SD = 4.23) to post-test (M = 19.06, SD = 6.37) showing a decrease in feelings of worry and the control group means decreased from pre-test (M = 23.79, SD = 4.02) to post-test (M = 22.57, SD = 5.46) showing a smaller decrease in feelings of worry.

The interaction between group and timepoint was not significant, F (1,29) = 1.89 p = .18, η^2_p = 0.06. (see Figure 8.13; Appendix H, Tables H.53-H. 56)

Figure 8.13

Worry Measured by the Adapted Thoughts' Subscale Mean Values for Students Identified as High in Test Anxiety



8.5 Research Questions That Arose from This Study

8.5.1 Research Question 4

As I administered the intervention to two schools, $n = 32$, I ran a 2 x 2, mixed, factorial ANOVA to check for any experimenter effect between the groups where I had administered the intervention against similar schools.

I ran the intervention with Year 6 in an independent, co-educational, all through school ($n = 12$) and in an independent all through girls' school ($n = 20$). To gain the best comparison, I ran the ANOVA comparing against an independent all through co-educational school (belonging to the same cluster of schools as my school, $n = 15$) and an independent all through girls' school, ($n = 15$).

The main effects of the ANOVA for the independent variable, Timepoint, $F\ (1, 60) = 3.52$, $p = .06$, $\eta^2_p = 0.05$ The effect was not significant.

The main effect was not significant for the independent variable Group, $F\ (1, 60) = 0.003$, $p = .95$, $\eta^2_p = < .001$. There was no significant interaction, $F\ (1,60) = 0.17$, $p = .68$, η^2_p

= 0.003. (see Appendix G, Tables G.50 – G.52) For descriptive statistics (see Appendix G, Table G.53)

8.5.2 Research Question 5

The result of the independent *t* test on the Adapted Thoughts' subscale for both pre and post timepoints was non-significant. The pre-test t (49) = .60, p = .55, Cohen's d = 0.17, post-test t(49) = .38, p = .71, Cohen's d = 0.11. The pre-test means for the high stakes' tests online group was M = 17.83 and the post-test mean for the online group was M = 17.37. For the group taking high stakes' paper tests the pre-test mean was M = 16.95 and the post-test mean was M = 16.81. (see Appendix G, Table G.56)

9. Discussion

This chapter interprets the results and makes links to the ideas discussed in the literature review. In addition, wider contributions to the theoretical fields of cognitive science, test anxiety and self-efficacy are considered. Limitations of the study are explored. As the strands of my thesis that are cognitive science, psychology and education come together, implications for classroom practice and education are considered as per the spirit of the EdD qualification. The intervention, further implications and possible educational reforms are reflected upon in Chapter 10.

I would like to highlight a key finding from my study and that is that the intervention was successful in increasing feelings of self-efficacy in test taking for the group of students, who are high in test anxiety. However, the results in this study are exploratory. The subscales that I adapted and created to measure the data have not been cross validated on a further sample and thus all results must be interpreted with that caveat in mind.

The following sections follow the research questions as they are key to the structure of this chapter.

9.1 Worry Reduces in Both the Intervention and Control Groups

In Research Question 1, I asked if the intervention that uses retrieval practice to teach the testing effect could reduce feelings of worry in 10-to11-year-olds. In the introduction to this thesis, test anxiety was highlighted as a problem for 10-to 11- year-olds and increased testing for this age-group was noted as a common factor (e.g., Harlen et al., 2002). Worry was identified as a key element in the test anxiety construct (Liebert & Morris, 1969). This identification led to my choice of measure in the Children's Test Anxiety Scale (Wren & Benson, 2004) as one key subscale, thoughts, focussed on the worry or cognitive element of test anxiety.

9.1.1 Interpretation of Results

The idea of worry and its roles in an individual's perceived control (Putwain & Aveyard, 2018) and in reducing uncertain control (Putwain & Pescod, 2018) in high stakes' tests were key to the design of the intervention that used retrieval practice of the testing effect to increase self-efficacy to try and reduce worry and increase control in evaluative situations.

Agarwal et al. (2014) suggest that retrieval practice made 72% of students feel less anxious about taking tests. The idea of using retrieval practice to improve individual control

in taking tests is key to this study, but the emphasis is different in my study because the retrieval is not of the subject content that is being tested on the exam but retrieval of a test taking strategy to build self-efficacy in test taking (alongside exam practice) and thereby, provide pastoral support to students.

The results from the mixed, 2 x 2, factorial ANOVA show that the worry element of the test anxiety construct, as identified by Liebert and Morris (1967) reduced from pre-test to post-test for both the intervention and the control group. The reduction was larger for the intervention group but there was a reduction of worry for both the intervention and the control group.

Perceived control (Putwain & Aveyard, 2018; Schunk & Pajares, 2005; Weiner, 2010) and uncertain control (Putwain & Pescod, 2018) are part of the self-knowledge beliefs strand from the S-REF (Self-referent executive function) model (Zeidner & Matthews, 2005) of test anxiety and were the areas of the test anxiety construct that were targeted by the intervention. The intervention was designed to increase self-efficacy in test taking and alter self-knowledge beliefs, thereby reducing worry, increasing perceived control, and reducing uncertain control. Potentially, the theories of the testing effect, test anxiety and self-efficacy combine in the intervention in a dynamic similar to that of the triad of reciprocal causation (Bandura, 1986b). I tried to influence (through the intervention) test taking behaviour through what individuals could draw from their environment and their personalities by making them feel more self-efficacious about taking tests.

Both the intervention and the control group experienced a reduction in worry at the post-test stage and this result may reflect the impact of the intervention for the intervention group. These results may suggest that the intervention altered the dynamic of the S-REF (Self-referent executive function) test anxiety model (Zeidner & Matthews, 2005) impacting the self-knowledge beliefs strand to create feelings of self-efficacy about taking tests. It is possible that the intervention provided the explicit and frequent feedback that Bandura (1997) recommends for students who lack self-efficacy. However, because there is a decrease in worry for both groups, it is important to consider why this result has happened and to acknowledge that reduction in worry or cognitive test anxiety may not be a direct result of the intervention.

It may be that the reduction in worry for both groups reflects changing trends in educational practice. At the time that I started my study in 2016, retrieval practice in

classrooms, to my knowledge, was not commonplace although the learningscientists.org (https://www.learningscientists.org) had been created by then. The inclusion of retrieval practice in the U.K. Initial Teacher Training Core Framework (2020) and the increasing dissemination of this practice through organisations such as retrievalpractice.org, (https://www.retrievalpractice.org), learningscientists.org (https://www.learningscientists.org) and the launch of the Chartered College for Teachers (https://chartered.college) whose journal IMPACT was dedicated to cognitive science in January 2020, the blog from the Effortful Educator (https://theeffortfuleducator.com) about retrieval practice featuring interviews with key researchers in the retrieval practice field, the books 'Make It Stick' (Brown et al., 2014) and 'Powerful Teaching' (Agarwal & Bain, 2019), may mean that students are practising testing more regularly in the classroom as a result of teachers learning about retrieval practice. A paper that highlights the benefits of testing in schools (Roediger, Agarwal et al., 2010) has 95 citations on Google and 2,940 views on Academia.Edu as of February 2021. In addition, Mike Hobbiss, (Hobbiss, 2021) tweeted "most useful single paper for teachers that I have ever read" with reference to Yang, Luo et al. (2021) meta-analyses about the testing effect.

These examples testify to the current popularity of ideas about retrieval practice and may explain why the control group also experienced a reduction in worry at the post-test measure that suggests a level of confidence in test-taking.

9.1.2 Links to Test Anxiety

The students in this study may have experienced much more retrieval practice in the form of low stakes tests in schools and may be aware of the benefits of the testing effect. Perceptions about testing that have historically been viewed as negative occurrences (e.g., Bjork et al., 2013; Dirkx et al., 2019) may be changing in schools.

The Adapted Thoughts' subscale was established through exploratory and confirmatory factor analysis and Question 28 'I feel scared' loaded three times out of eight onto the thoughts' subscale instead of the autonomic reactions' subscale that was its loading on the original scale.

The history of test anxiety has many models and constructs but the Liebert and Morris construct (1967) of worry (cognition) and emotionality (behaviour) was an important step in the definition of the construct. Wren and Benson (2004) themselves suggest that they may have defined the test anxiety construct too narrowly with their three subscales (*thoughts*

representing worry, *autonomic reactions* representing emotions and *off task behaviours*) and recommend that future studies look at alternative test anxiety content for children. The transactional model of test anxiety, the S-REF (Self-referent executive function) model (Zeidner & Matthews, 2005) that was central to this study, includes alternative content because it is a dynamic model drawing from the individual's beliefs and circumstances. Moreover, Putwain et al. (2020) suggest that test anxiety is a multidimensional construct. The cross loading of Question 28, 'I feel scared' onto the thoughts' subscale may suggest that 'feeling scared' is being experienced by students as a worry, a cognitive response, more so than a feeling, for the data in this study.

The loadings of the items for the Adapted Thoughts' scale in the exploratory factor analysis suggest that the test anxiety construct is dynamic and complex as there are cross loadings across the subscales and it is possible that different students experience the different items on the questionnaire sometimes as thoughts, sometimes as feelings (autonomic reactions) and at different times in the testing process. The S-REF (Self-referent executive function) model (Zeidner & Matthews, 2005) incorporates this transactional element of test anxiety and the Adapted Thoughts' subscale is an attempt at integrating children's test anxiety with alternative content to represent the more dynamic nature of transactional test anxiety. My attempt uses a similar principle to Putwain et al. (2020), von der Embse et al. (2021) in treating test anxiety as a multidimensional construct. However, in my study, the research is focussed on the cognitive element of test anxiety. The exploratory factor analysis of the Children's Test Anxiety Scale allowed me to see if any items from the other subscales loaded onto the cognitive or thoughts' subscale in my study and allowed me to explore the multidimensional nature of test anxiety before I adapted the established measure.

The Adapted Thoughts' subscale consists of eight items, to measure the cognitive test anxiety element of worry in children and was created by me, a teacher / researcher and validated through Cronbach's *alpha*, confirmatory and exploratory factor analysis as well as measurement invariance and may add to the few existing measures for test anxiety in children. The brevity of the subscale may be an asset for use with 10-to 11-year-olds and the Adapted Thoughts' subscale could easily be combined with the Self-efficacy in Test Taking subscales to form a type of multidimensional test anxiety measure that is still only 21 items in total. It would need cross validating in a further sample.

9.1.3 The Multilevel Linear Model

The problem with both the latent means analysis and the 2 x 2 mixed, factorial ANOVA is that the potential hierarchical (classroom / school) nature of the data is not considered. The multilevel linear model allowed me to look at the data at classroom / school level as well as at participant level. Usually, an ANCOVA can be run to control for pre-test scores and provide a more informative analysis (e,g., Dugard & Todman, 1995) but because there was no homogeneity of regression slopes, the assumption that must be met for ANCOVA analyses (Field, 2012), I used the multilevel linear model.

The scatter plots (Figure 8.2 and Figure 8.3) of the intervention and control groups' data and for the different schools showed that there were clusters of pre-test and post-test scores and the multilevel linear model has allowed me to see whether these clusters were happening in schools or at the individual participant level. Given the difficult but essential (e.g., Nelson & Narens, 1994) nature of conducting interventions in the real-world environment of classrooms, this information is useful.

An important aspect of the multilevel linear model is the Intraclass Correlation that shows the total variability in the outcome (post-test scores) that can be attributed to the different classes. The Intraclass Correlation is a way of showing how the classes behave. I think of it as a group dynamic measure. A large Intraclass Correlation suggests that there is little variability in the outcome within the class (the students are doing and thinking the same things in their classes). The opposite is true. If the group dynamic in the class has little effect on the individuals, then the outcome (post-test score) varies greatly within the class and there is a small Intraclass Correlation (Field, 2012).

The Intraclass Correlation denotes whether there is significant clustering in an observation (Heck et al., 2014). For the first model, the variance components model, the Intraclass Correlation is 11 % for the Adapted Thoughts' subscale. This result suggests the proportion of variation in post-test scores that lies between the schools is 11%.

When the multilevel linear model was changed to include the explanatory variables of groups and the centred grand means of the pre-test scores as level 1 fixed effects with schools as level 2 random effects, the random intercept model, then the Intraclass Correlation reduced to 5%.

Looking at the data in this way allowed me to capture a more natural picture of what is happening (Buchanan, 2017) because not all students start in the same place (fixed

intercepts). Looking at random effects' models rather than fixed effects' models means that various sources of variance can be considered (e.g., Crowson, 2019; Field, 2012; Heck et al., 2014). Fitting the random intercepts model with level 1 predictors of group and pre-test scores resulted in a significant $\boldsymbol{\Delta}X^2$ from the variance components' model. This result suggests that the pre-test scores that are in clusters are affecting the post-test score outcomes but that this result is mainly at the participant level. The clustering of pre-test / post-test scores in the data may represent a limitation in terms of the very nature of using questionnaires with a Likert type four-point scale. (see Limitations).

The Intraclass Correlation suggests that 11% of the sample's variation in outcome (post-test scores) is attributable to the school or class the participants were in. This result means that for 89% of the sample, the participants' outcomes (post-test scores) are varied within the class. In other words, it is not specific schools that are behaving very differently to the majority. However, this result does suggest that approximately one school (11%) is contributing to the clustering.

Many of the clustered data points are from the control group's schools. I examined 57 data points in the clusters and 38 points were from control schools. Thus, one can conclude that the control group's schools are more likely to be contributing to the clusters of pre-test and post-test scores than the intervention group's schools.

As a teacher and researcher, this result suggests to me, that the students, in the control group are generally feeling confident about taking tests at the pre-test and post-test stage and that confidence in test taking is occurring at the participant level rather than at the school level, apart from one school that does account for 11% of the clustered scores. Therefore, the multilevel linear model may suggest that there is confidence in test taking at the individual participant level especially for the control group. It is possible that the school that took online tests rather than actual high stakes exams' is contributing to the level of test taking confidence. This school is identifiable in the centre of the clusters. An independent *t* test was run (Research Question 5) to see if there was a difference between this school and another similar school and there was not a significant difference.

9.1.3.1 Limitations of the Multilevel Linear Model. The sample size for the level two units, the schools, may be a limitation of the model. There is no real rule of thumb measure (Field, 2012) but it is commonly accepted that the larger the better (Field, 2012). In addition, I used grand centering of the mean for the level one fixed effect predictor, pre-test

score because I wanted to focus on the level two variable of school and control for the level one covariate of group and pre-test scores (Field, 2012). However, Antonakis et al. (2019) recommend using a cluster mean to ensure the assumptions for the random effects model are met.

9.2 Self-efficacy in Test Taking Increased in the Intervention and the Control Group

In Research Question 2a, I asked if the intervention increased feelings of self-efficacy in test taking in 10-to-11-year-olds. In the literature review, I suggest that by manipulating feelings of self-efficacy in test taking at the self-knowledge beliefs strand in the S-REF (Self-referent executive function) model of test anxiety, (Zeidner & Matthews, 2005) where maladaptive coping may set in, may serve to reset the appraisal of test taking self-efficacy ahead of the evaluative situation itself (Schutz et al., 2014). The findings are interpreted here for the Self-efficacy in Test Taking and Little Self-efficacy in Test Taking subscales that were used as measures for this question.

9.2.1 Interpretation of Results

Both subscales were created using exploratory and confirmatory factor analysis as well as tested for measurement invariance across gender, group and timepoint. The results from the mixed 2 x 2 factorial ANOVA for the Self-efficacy in Test Taking subscale (where higher scores represent higher self-efficacy in test taking) show that feelings of self-efficacy in test taking increased for both the intervention and the control group. The increase was slightly larger for the intervention group.

The results from the mixed 2 x 2 factorial ANOVA for the Little Self-efficacy in Test Taking subscale (where lower scores represent higher self-efficacy in test taking) show that feelings of self-efficacy in test taking increased for the intervention and control groups. There was a significant interaction between group and timepoint suggesting that students in the intervention group benefited significantly from the intervention and experienced an increase in feelings of self-efficacy in test taking when measured with the smaller subscale.

Both the intervention and control group's self-efficacy in test taking increased from pre to post-test when measured with both of the subscales. It is possible that the intervention manipulated the self-knowledge beliefs strand of the S-REF (Self-referent executive function) model (Zeidner & Matthews, 2005) of test anxiety to make students feel more self-efficacious in test taking. It may be that students, as a result of the intervention, feel more self-efficacy for test success (Smith et al., 1990). It is possible that the intervention has

worked to intercede at the forethought level when test anxiety manifests (Schutz & Davis, 2000; Schutz et al., 2014) to reduce a perceived threat and give more control (Putwain & Aveyard, 2018; Putwain & Pescod, 2018). However, because self-efficacy in test taking increased for both the intervention and control group when measured with the self-efficacy in test taking subscales, it is also possible that students may generally be feeling more confident about taking tests because it is becoming more common to practise tests in class as teachers are more familiar with the benefits of retrieval practice.

There was a significant increase of feelings in self-efficacy for the intervention group when measured with the Little Self-efficacy in Test Taking subscale. It is conceivable that the negatively worded items on this subscale elicited responses more easily from the students. This outcome may be because the language in the questions is more emotive and resonated more with students, especially students who are high in test anxiety and who may have been in a more emotive state when completing the questionnaires (e.g., Sarason, 1972; Schutz et al., 2002). The results may suggest that the intervention worked very well with students who identify with more negative feelings about test taking than positive ones. Williams et al. (2001) propose that negatively worded items performed better than positively worded items with middle school students; a finding that contradicted Benson & Hocevar (1985) where elementary school students were less likely to show agreement through disagreeing with negatively worded items than showing agreement for positively worded items. Responses to wording of items on a questionnaire is another variable that needs to be considered as a potential limitation when using questionnaires (e.g., Weems et al., 2003).

Perhaps, a clue to these responses may lie in the *fear appeals* research (e.g., Putwain & Best, 2011; Putwain et al., 2014). For students who are low in test anxiety, fear appeals (reminders about upcoming high stakes' tests and their possible consequences) can work to motivate some students but for students who are high in test anxiety, fear appeals can make them more anxious. It is possible that the negatively worded items on the Little Self-efficacy subscale had more impact with students who are high in test anxiety and at various levels approaching a high level.

The ANCOVA for the Little Self-efficacy in Test Taking Subscale that adjusted for the covariate pre-test scores on the dependent variable of the post-test score for the intervention and control group showed a decrease in the marginal means from the descriptive means for the intervention group. This result suggests an increase in self-efficacy in test

taking for the intervention group. For the control group, the marginal means increased suggesting a decrease in self-efficacy in test taking for the control group.

This result may suggest that the intervention worked to increase self-efficacy in test taking with the intervention group. It may suggest that the control group would have benefited from the intervention because self-efficacy in test taking reduced. However, the problems applying to the completion of pre-test and post-test questionnaires and Likert scales (see 9.2.3) may apply. In addition, the Little Self-efficacy in Test Taking subscale has not been cross validated and so all results are exploratory.

9.2.2 Links to Control

In terms of control and test taking, Schunk and Pajares (2005) suggest that people expect success relative to the measure that they feel in control of their behaviour. The increase in self-efficacy in test taking could be as a result of the intervention for the intervention group who feel more in control. Knowing that their 'testing routes are well oiled' that was a key phrase in the intervention, borne from the idea of recall facilitating recall, the idea of memory as a modifier (Bjork., 1975; Bjork, 2012) may have increased their perceived control (Pekrun, 2006; Putwain & Aveyard, 2018) and reduced uncertain control (Putwain & Pescod, 2018) by altering the strand of self-knowledge beliefs in the S-REF (self-referent executive function) model of test anxiety (Zeidner & Matthews, 2005)

It is possible that the intervention manipulated how the students felt about taking tests and altered a social learning model of test anxiety (Smith et al., 1990) to improve any potential loss of self-control. Potentially, for the control group, any change to their perceived and uncertain control may be a result of more retrieval practice with low stakes tests. Both groups were taking practice tests at the time of the intervention and thus these practice tests may have contributed towards feelings of self-efficacy in test taking and therefore control. This variable has not been measured although it has been commented on in the teacher and student feedback.

Smith et al. (1990) suggest that self-efficacy for test taking success as part of the social learning model contributes towards test anxiety and academic performance as well as cognitive attentional models of test anxiety. If the results of Research Question 2a are regarded in the light of this statement, then the interaction that shows a significant increase in feelings of self-efficacy in test taking for the intervention group when measured with the Little Self-efficacy in Test Taking scale would support the suggestion from Smith et al.

(1990) that interventions that are multimodal are important. The caveat that needs to be applied to my study is that the effect of taking practice tests alongside the intervention has not been measured as a separate process.

The findings from this study may challenge the idea proposed by Putwain & Aveyard (2018) that younger children and girls have less perceived control than older children. This result may be because of increased use of retrieval practice in classrooms or as a result of the intervention to increase self-efficacy in test taking and thereby improve perceived control. The findings may support Agarwal et al. (2014) in that retrieval practice provides a strategy to reduce test anxiety. To my knowledge, retrieval practice of the testing effect itself has not been used before as a test anxiety intervention.

Furthermore, and related to the literature around fear appeals (e.g., Putwain & Best, 2011; Putwain et al., 2014) and the role of the teacher in test anxiety (e.g., Schutz et al., 2014), Putwain et al. (2014) suggest that students respond in different ways to fear appeals and it may be that for students, who are high in test anxiety, the reassurance of hearing from a teacher that 'your testing routes are well-oiled' is a really important phrase and the results are shown by the significant result in the interaction for this research question when measured with the Little Self-efficacy subscale.

The S-REF (Self-referent executive function) model (Zeidner & Matthews, 2005) was the key model of test anxiety identified in the literature review. I designed the intervention to target the self-knowledge beliefs' strand of this model with the aim of moderating test anxiety at this stage of the test anxiety process by giving students some control over feelings about test taking.

The research about perceived control (Putwain & Symes, 2018) and uncertain control (Putwain & Pescod, 2018) recommend that interventions that are designed to alleviate the worry or cognitive element of test anxiety are unlikely to improve performance unless they also target control (Putwain & Symes, 2018). They state that there is no empirical evidence for how self-belief is acting with the cognitive element of test anxiety to predict exam performance. Zeidner (2014) proposes a crossover is needed between test anxiety and other theories. The intervention and creation of a self-efficacy in test taking measure provided this cross over of theories and the exploratory results from this study offer some empirical evidence about how self-belief (self-efficacy in test taking) is acting with cognitive test anxiety although performance is not measured in this study.

9.2.3 Limitations to Research Questions 1 and 2a

I do not know if just taking practice tests which is what all the schools were doing in the run up to high stakes' tests are contributing towards a confidence level in test taking. This variable has not been measured. Teacher feedback (see Appendix D) suggested that a combination of learning about the testing effect and taking practice tests works.

Figure 8.2 and figure 8.3 that show clustering on the Adapted Thoughts' subscale that was used to measure cognitive test anxiety or worry may suggest that the students want to appear confident or are not sure what to fill in on the questionnaires, so aim for the 'some of the time' options. This problem may arise from using pre-test and post-test questionnaires with Likert type scales.

In addition, D'Mello et al. (2013) discuss learners' being more likely to under report positive experiences. Kivikangas et al. (2011, p.183) refer to self-report items being subject to "participant answering style, social desirability, interpretation of questionnaire item wording, or limits of participant memory'. My experience as a teacher suggests that the age-group in this study tend to be truthful about their feelings but I acknowledge that students can be susceptible to peer, parental and teacher pressure and all students are different. D'Mello et al. (2014) refer to learners' honesty and willingness to report emotions and social pressure when reporting negative emotions and accuracy at being able to label one's own emotions. It is possible that the students were unsure of their emotions at either pre-test or post-test. It is possible that they were dishonest. It is also possible that some responses are inaccurate because the right answer was not available as an option for a participant. Variations and combinations of the above ideas may also exist for individuals.

Moreover, it is also possible that the very nature of a four-point Likert type scale was not the best choice for a questionnaire as it may have encouraged students to opt for the middle ground and that this option is what we can see in terms of the clusters of the data for the Adapted Thoughts' subscale. However, the Children's Test Anxiety questionnaire does have good validity and has been used repeatedly with children of this age-group (e.g., Aydin & Bulgan, 2017) and the Self-efficacy in Test Taking questionnaire was designed with this aspect in mind, to mirror the Likert type scale to make it easier for 10-to11-year-olds.

Sarnoff et al. (1958) suggest that some students may be reluctant, possibly through societal conditioning, to acknowledge their own feelings in questionnaires. As a teacher, I have observed a sense of debrief when students come out of an exam situation; feelings of

relief and / or worry about the outcome are expressed enthusiastically and it is possible that these feelings may affect the emotive state at the post-test stage and consequently, the post-test score. As a teacher, who has witnessed many students leaving exam halls, this scenario is the 'What did you get for?' 'I only wrote', 'I did not write that', 'Oh no, I didn't see that question' type of communication that students exchange in when leaving the exam room. I have seen students cry as they leave exam halls when they realise that they have answered incorrectly or missed a question out.

In addition, there is a potential problem because of the timing of the post-tests. The post-test timepoint was a variable that was difficult to control for in this real-world research setting and some high stakes' tests were conducted over a period of four weeks so inevitably the timings of the post-tests varied from school to school. The pre-test / post-test design may have worked in enabling me to measure change, but it is also possible that just by completing a pre-test that the groups' responses at post-test have been altered (e.g., Shuttleworth, 2009)

Sommer and Arendasy (2014) discuss the effect of timepoint on measurements for trait test anxiety in their paper that examines measurement bias and test anxiety. In their study, the post-test questionnaire resulted in a high, test anxiety to performance correlation. This result may be because students are anxious about the way they have just performed on the test. Zeidner (1998) suggests that the levels of state anxiety that are experienced by students during tests may prime emotions remembered from previous tests. It is possible that answers in post-test questionnaires may be affected by recent and past test experiences.

These potential issues are compounded by test anxiety being a difficult construct to measure because of the reliance on pre and post-tests and the inability to be able to measure test anxiety at the very beginning or the middle of a test because such a measure could affect performance in a test (e.g., Schutz et al., 2014). An interruption of this kind could be disastrous for an individual candidate and exam regulations would be broken. There is the notable exception of Galassi et al. (1981). It is possible that the timing of the post-tests has contributed to the results in my research questions. The self-efficacy questionnaires were administered alongside the Children's Test Anxiety Scale questionnaires and thus the arguments outlined by Zeidner (1998), D'Mello et al. (2014) and Sarnoff et al. (1958) apply to both Research Questions 1 and 2a.

9.2.3.1 Limitations of the Self-efficacy in Test Taking Questionnaire. The design of the Self-efficacy in Test Taking questionnaire may be a limitation in this study. Using a

Likert type scale without a midpoint and modelled on the Children's Test Anxiety Scale Likert type responses was, in my opinion, probably a good idea as these options mean that the upper primary aged students could not simply opt for the middle response value and the two questionnaires were similar and therefore, potentially more accessible. There may be some issues around the negatively worded items on the questionnaire. This wording was formulated to avoid response bias in the answers (Chyung et al., 2018) but a lot of researchers are now critical of this practice (e.g., Benson & Hocevar 1985; Chyung et al., 2018). In defence of this process, the negative items were kept to a minimum and mainly grouped together in the questionnaire and have been interpreted separately.

From my experience as a teacher of this age-group, I think it is possible that the negative wording of the items on the Little Self-efficacy Scale in Test Taking subscale 'panic', 'negative thoughts', 'doubt' had more impact on the students in terms of their automatic responses and that although the negatively worded questions weren't signposted as such (Chyung et al., 2018), the negative emotive language is unlikely to have been misinterpreted and may well have served as a 'wake up' call should a participant be feeling fatigued. (e.g., van Sonderen et al., 2013) Or, perhaps, I must consider that the students were confused by the questions and answered in the same way as the larger subscale. (Chyung et al., 2018)

It may be that reverse worded items confused the students (e.g., Weems et al., 2003; van Sonderen et al., 2013). However, reverse worded items are often recommended as good practice in questionnaire design (e.g., Swain et al., 2008). Exploratory factor analysis and confirmatory factor analysis confirmed two subscales for the items in the questionnaire and two items were removed (Q5 'When I take tests, I feel anxious about being in a different room to my usual classroom' and Q6 'When I take tests, I feel anxious about the instructor' to create the Little Self-efficacy in Test Taking subscale.

The data were measured with both subscales and supports the ideas expressed by Hartley (2013) who suggests that reverse worded coded items' results should be presented separately; obtaining two subscales and reporting on both of these managed this process effectively.

Using positively and negatively worded items in the questionnaire may guard against response bias such as acquiescence bias, (Chyung et al., 2018) or inattention and confusion (van Sonderen et al., 2013). One way to guard against fatigue is to create a short

questionnaire and to group negatively worded items together. There were only four items out of fifteen that were negatively worded, and they were partially grouped together. Using a four-point scale is a good idea according to Garland (1991) as it forces the respondent to make a choice. One can improve the quality of data and reporting by completing all analyses separately (Hartley, 2013).

Benson and Hocevar (1985) discuss the impact of item phrasing with elementary students. The design of the Self-efficacy in Test Taking questionnaire could have been improved through seeking children's input. The Self-efficacy in Test Taking questionnaire was not used in the pilot but it was trialled by me in a classroom before other teachers used it in the main study. My Year 6 students, who were a mixed ability class, accessed the questions readily but I hadn't consulted them in the creation of the questionnaire. In terms of seeking expertise (e.g., Putwain, 2020) in the design of the questionnaire, I presented the proposed Self-efficacy in Test Taking questionnaire to the INSTRUCT (Implementing New Student Thinking Resources Using Cognitive Theory) lab and received feedback about the wording of the items on the questionnaire.

I managed the self-efficacy data by creating two subscales using the robust processes of exploratory and confirmatory factor analyses and measurement invariance. They were tested by Cronbach's *alpha* for reliability and fit statistics confirmed the subscales. Therefore, I have approached the analyses of data from a thorough perspective (e.g., Zhang 2016).

Although the results from the Self-efficacy in Test Taking subscales are exploratory because both subscales need cross validating on another sample, they allow me to explore the fundamental relationship that is the role of self-efficacy in test taking to the element of worry in the test anxiety construct and this relationship is the basis of Research Question 2b.

9.2.4 Implications for the Classroom from Research Questions 1 and 2a

In the spirit of the EdD qualification, the results of this study that incorporate the theories of cognitive science, specifically, the testing effect, test anxiety and self-efficacy are presented here in terms of implications for education, for use in the classroom. In Chapter 10, there is an evaluation of the intervention itself and further reflections and thoughts for curricula change and policy reform.

Putwain, Chamberlaine et al. (2014) report on the success of the STEPS (Strategies to Tackle Exam Pressure and Stress) intervention for 15- to16 -year- olds that was successful in

reducing test anxiety in students with high test anxiety. One aspect of the intervention was replacing negative self-talk with positive self-talk.

The phrase ‘your testing routes are well-oiled’ that was key to the intervention and was my way of interpreting the idea of memory being a modifier (Bjork, 1975) as recall promotes recall, does encourage positive self-talk /thinking. I wanted the students to feel that their practice testing had prepared them well for tests and to know that the reason why the testing effect works. It was explaining why the testing effect works that was important.

The results for the first two research questions suggest a reduction in worry and an increase in self-efficacy for both the control and intervention groups. It is possible that the intervention worked for the intervention group and potentially the writing tasks in the intervention help to manage test anxiety. These were not measured separately or specifically but constitute a third of the intervention. These findings may add to the literature that suggests writing about feelings about testing can reduce stress in high stakes’ tests. (e.g., McKeachie, 1951; Ramirez & Beilock, 2011) Especially, writing about testing and self-efficacy can reduce test anxiety (Lang & Lang, 2010). In this study, the retrieval of a high stakes’ testing strategy features two weeks of writing. A letter to Bob Bjork about the testing effect and also a letter to a friend giving advice about how to manage test anxiety. (see Appendix D). Teachers can use the intervention or perhaps aspects of the intervention, such as the writing tasks, to boost their classes’ *cognitive morale* prior to high stakes’ testing. Furthermore, learning about the testing effect will enable students to understand why they are being tested. It may enable students to make more accurate judgements about what they know and need to learn. Knowing how memory works is, in my opinion, a cognitive advantage, for life.

The majority of student and teacher responses support the idea that learning about the testing effect through retrieval practice helps them feel more self-confident when taking tests, especially when combined with practice tests. (see Appendix D) It gives the ‘why’. Giving students the ‘why’ means giving them control. (Martin 2002, 2007; Putwain, 2018) This idea supports Zepeda et al. (2020) and their proposals for more research into motivational links and desirable difficulties and also supports Agarwal et al. (2014) findings about how retrieval practice reduces test anxiety with this age-group.

Moreover, the results suggest that the intervention worked particularly well with students, who experience high test anxiety, an average of 10-20% in any classroom. This

number is significant and may give these students a coping strategy to allow them to control negative thinking. The discussion for Research Question 3 expands on this idea.

The implications for schools are that this intervention can be delivered by a classroom teacher and teaches students about the testing effect, making the link for students about why testing works and may act as a sort of silver psychological educational bullet (Yeager & Walton, 2011) to increase feelings of self-efficacy in test taking, especially for students who are high in test anxiety.

On a personal note, when I was discussing this study at a Girls' Schools Association training occasion, one of the teachers came up to me and said: 'That's the missing link; the missing piece of the jigsaw puzzle; teaching the children how testing works on the brain'. Teaching students the cognitive theory of retrieval practice using the testing effect may be the missing link in creating feelings of self-efficacy in test taking and giving students more belief so that worry is reduced.

The new Adapted Thoughts' subscale and Self-efficacy in Test Taking subscales have good reliability and validity and were created by a teacher and used in schools and can be used in classrooms once they have been cross validated.

9.3 MIMIC Models and the Relationship Between Test Anxiety and Self-efficacy in Test Taking.

In Research Question 2b, I asked what self-efficacy in test taking sum scores tell us about individual test anxiety and have used MIMIC (Multiple Indicators, Multiple Causes) models to explore this relationship.

9.3.1 Interpretation of results

MIMIC (Multiple Indicators, Multiple Causes) models are a way of using Confirmatory Factor Analysis with a covariate to examine the effect of an explanatory variable on a latent variable (Finch & French, 2015). An idea posited in the literature review is about developing more control through reducing uncertain control (Putwain & Pescod, 2018) and improving perceived control (Putwain & Aveyard, 2018) in evaluative situations. By examining the specific relationship of self-efficacy in test taking on worry, the cognitive element of test anxiety (Liebert & Morris, 1967) I hope that the results contribute some ideas as to how worry and self-efficacy in test taking relate to each other and how this relationship may contribute towards increased control in test taking.

The MIMIC (Multiple Indicators, Multiple Causes) models have allowed me to examine the significance of the relationship between self-efficacy in test taking and worry as measured by the latent variable from the confirmatory factor analysis and subsequent measurement invariance of the Adapted Thoughts' subscale.

I ran MIMIC (Multiple Indicators, Multiple Causes) models for intervention and control groups at pre-test and post-test conditions with both the Self-efficacy in Test Taking subscale and the Little Self-efficacy in Test Taking subscale to look at the relationship between self-efficacy in test taking sum scores and the worry element of test anxiety. The self-efficacy in test taking subscale sum scores were the regressed covariate to the latent model of adapted thoughts.

If the estimate is statistically significant then we can conclude that the covariate is related to the factor (Finch & French 2015). In this case, if the estimate is statistically significant, we can conclude that feelings of high /low self-efficacy in test taking are related to the level of worry about test taking. Where the covariate is continuous as it is in this case, (sum scores in the Self-efficacy in Test Taking Subscales) then β, the estimate, can be interpreted just like the slope in a standard regression analysis. (Finch & French, 2015).

Results of the MIMIC models (Multiple Indicator, Multiple Causes) models (see Appendix G, Table G.57, Figures G.17 – G.28) show that for all of the intervention groups, at pre-test and post-test, and with both subscales that the high and low self-efficacy is significantly related to worry and that there is an inverse relationship: high worry and low self-efficacy, high self-efficacy and low worry. For the control group, the *p* values between high and low self-efficacy and worry do not show a difference. This result suggests that there is no relationship between self-efficacy in test taking and worry for the control group. This result may reflect the clusters of data, particularly for the control group in the Adapted Thoughts' data.

It is possible that the intervention group were more conscious of their feelings about test taking as a result of receiving the information from the intervention and that this feeling is represented in a more wide-ranging data that gives the significant inverse relationship seen in the multiple indicator and multiple causes model for the intervention groups. Path diagrams show the relationship from the observed variable, self-efficacy in test taking (measured with both subscales) to the factor, the score on the Adapted Thoughts' subscale. (see Appendix G, Table G.57, Figures G.17 - G.28)

9.3.2 Limitations

Opposing views exist in the literature about the use of MIMIC models. Lee et al. (2013) suggest that formative MIMIC models should not be used as they do not provide a valid method of integrating formative variables into empirical studies and conceptualizing a causal relationship is illogical. On the other hand, Diemer (2019) wonders why MIMIC models are not used more commonly. In this study, I did not seek to establish a causal relationship between self-efficacy in test taking and worry but have used the MIMIC model simply for exploration of the relationship between the self-efficacy data to the Adapted Thoughts' data by looking at the regression of the covariate to the latent variable for the intervention and the control group. I followed the procedure for calculation and interpretation from Finch & French (2015) and was able to discuss my model and the continuous covariate used (personal communication, Holmes Finch, March 2021)

9.3.3 Implications

The results may add to the literature around perceived control being a critical factor in understanding the relationship between test anxiety and test performance (Putwain & Aveyard, 2018) and the reduction of uncertain control (Putwain & Pescod, 2018) being key to reducing test anxiety and is an example of trying to target one aspect (Putwain, 2018), the self- knowledge beliefs strand of the S-REF (Self-referent executive function) model (Zeidner & Matthews 2005) to reduce test anxiety.

It is interesting that the relationship between high and low self-efficacy to worry is only significant in all the conditions of the intervention group and not in any conditions of the control group. The Adapted Thoughts' data had clusters between response values one and two at pre-test and post-test especially for the control group and the intervention group's data is much more spread out. This result could represent the effects of the intervention or simply the participants themselves. The subscales need to be cross validated with another sample and these results are exploratory, but they may suggest that receiving an intervention designed to reduce worry or the cognitive element of test anxiety can alter feelings of self-efficacy in test taking. For this study and this data, the results suggest that an inverse relationship exists between feelings of high and low self-efficacy in test taking and low and high worry respectively for the intervention group.

9.4 Students Who Are High in Test Anxiety Benefit Significantly from the Cognitive Intervention.

In Research Question 3, I asked if feelings of worry can predict how well student self-efficacy changes in the intervention group? This question is answered by analysing the data relating to students who are high in test anxiety. In every classroom, there will be some students who are high in test anxiety. According to Cizek and Burg (2006), it is impossible to know how many are in a school. Putwain and Daly (2014) suggest a value of between 10 - 20% and in this study 18% of students were high in test anxiety ($N = 31$).

9.4.1 Interpretation of Results

The results suggest that the intervention of teaching students about the testing effect for students, who are high in test anxiety, increased their feelings of self-efficacy in test-taking using both the Self-efficacy in Test Taking and the Little Self-efficacy Scale in Test Taking subscales. Additionally, when measured with the Adapted Thoughts' subscale to look at worry, the intervention group's students, who were high in test anxiety, experienced a much larger reduction in their means than the control group, again suggesting that students, who are high in test anxiety, benefited from the intervention.

It may be that the intervention provided a coping strategy, intervening in the negative, irrational thoughts that can beset students who are high in test anxiety (e.g., Hembree, 1988; Zeidner, 2014). It is possible that the intervention enabled the students to keep the exam in perspective (Schutz et al., 2002). Students, who are high in test anxiety suffer with worrisome thoughts (e.g., Goonan, 2003; Sarason, 1972) and intrusive worry (Covington, 1982). For students who are high in test anxiety, the anxiety process can start with accessing negative self-beliefs (e.g., Matthew et al., 1999) so it is possible that the intervention intervened at this point of the S-REF (Self-referent executive function) model of test anxiety (Zeidner & Matthews, 2005) and the self-knowledge beliefs layer to increase self-efficacy and reduce negative self-beliefs and therefore reduce worry for students who are high in test anxiety.

Schutz et al. (2002) suggest that even learning test taking strategies can make students, who are high in test anxiety, anxious. In the control group, there was a reduction in self-efficacy in test taking for students, who are high in test anxiety when measured with the Little Self-efficacy scale and this result may link to the idea proposed by Schutz et al. (2002) in that even knowing that they need to study for an exam can be a trigger for anxiety and these students are often critical of themselves and their exam performance (Zeidner, 2014).

However, the results from the intervention group challenge the idea of learning about test taking making students more anxious (Shutz et al., 2002) as the benefits of the intervention when measured with both subscales in the Self-efficacy in Test Taking instrument show an increase in feelings of self-efficacy in test taking for these students. It is possible that the students in the control group would benefit greatly from the intervention of learning about the testing effect as the results suggest the intervention had a significant effect for students who are high in test anxiety when measured with the Little Self-efficacy in Test Taking subscale.

Young et al. (2003) refer to treating test anxiety through modifying the schema. In other words, small changes to the way that thoughts are linked. The results for the students who are high in test anxiety suggest that their schema may have been altered by the intervention. It is possible that they were able to access more self-efficacy in test taking thoughts through drawing upon their knowledge of the benefits of the testing effect and knowing they had practised this recall.

The idea of writing worrisome thoughts down as a coping strategy for test anxiety is not new (e.g., McKeachie, 1958). Lang and Lang (2010) suggest that competency priming ahead of an exam is beneficial for students who are high in test anxiety and improves test taking self-efficacy. Shen, Lang et al. (2018) suggested that positive writing ahead of taking a test benefited Chinese students. Ramirez and Beilock (2011) found that writing about testing helped students who were high in test anxiety.

There were two writing tasks in the intervention in week four and week six and the recall of the testing effect was needed to complete the tasks. The act of recall about the testing effect through writing a ‘Letter to Bob’ and then an advice letter to a friend, may have helped the students who are high in test anxiety feel more self-efficacious about test taking.

In this study, the intervention was delivered to all students in the class. This way of delivering the intervention was for two reasons: the potential positive effects on general well-being that a test anxiety intervention may yield (Weems et al., 2015) and, in my opinion, it is ethically apt with this young age-group. The problem with delivering an intervention to students, who are identified as high in test anxiety, and not to other students means that they could be subject to unwelcome peer observation and it may make them feel more anxious about the tests if they are treated differently to other students in the class.

9.4.2 Links to High Test Anxiety

This finding supports and adds to the literature about the success of interventions for students, who are high in test anxiety (e.g., Putwain, Chamberlaine, 2014.) There is a now new measure available that has good validity for its subscales, to measure a specific facet of the self-knowledge beliefs strand of the S-REF (Self-referent executive function) model of test anxiety (Zeidner, & Matthews, 2005); the Self-efficacy in Test Taking instrument. A key aspect to consider is that the intervention in this study may fill the gap identified recently by Putwain (2020, p.6) as a "pressing need" for test anxiety interventions for this age-group in particular, due to the high stakes' tests that Year 6 must sit.

Furthermore, the nature of retrieval practice has been shown to reduce test anxiety (Agarwal et al., 2014) but retrieval practice of cognitive theory to reduce test anxiety has not been tried before and the results suggest that retrieval practice of the testing effect is effective at increasing feelings of self-efficacy in test taking for students who suffer with high test anxiety. The potential wider impact of the results from this study for the classroom / schools are that in using retrieval practice to recall test taking strategies, in an intervention that teachers can deliver, can result in individual improvements in self-efficacy in test taking and that belief in capability, no matter what the ability leads to improved performance (Bandura, 1997). Moreover, a teacher /school may be able to provide a coping mechanism for students who are high in test anxiety to help in evaluative situations.

9.4.3 Limitations

One limitation is that although there is scale reliability and good fit statistics for the Self-efficacy in Test Taking instrument, the subscales have not been cross validated with another sample. Further limitations about the use of questionnaires to measure psychological characteristics have already been discussed. I think the nature of the wording on the Little Self-efficacy in Test Taking scale that was potentially more emotive that other questions may have resonated more with students who are high in test anxiety (e.g., Benson & Hocevar; Weems et al., 2003; 1985 William et al., 2001) but I have not measured this idea in any way in this study and think that more research is needed about how students, who are anxious about tests, respond to wordings on test anxiety questionnaires, especially for this age-group.

9.4.4 Implications for the Classroom

The results for the students who are high in test anxiety suggest that they benefited significantly from the intervention. The control group's self-efficacy in test taking reduced when measured at post-test with the Little Self-efficacy in Test Taking scale. This result may

reflect the nature of students who are high in test anxiety in being critical of themselves after an exam (e.g., Zeidner, 2014).

For teachers and educators, in every class there will be between 10-20% of students who are test anxious (Putwain & Daly, 2014). Putwain et al. (2020) and von der Embse et al. (2021) have developed, and cross validated a new scale with which to measure test anxiety for slightly older students. The aim is to identify these students to afford them the correct access arrangements that they need for exams such as extra time. In addition, Ofqual (Office of qualifications and exam regulations) published a recent report (Howard, 2020) acknowledging the problem of students who are high in test anxiety and discussing possible interventions. Writing about tests as a strategy to manage test anxiety is referred to as the report cites Ramirez and Beilock (2011). The Multidimensional Test Anxiety Scale (Putwain et al., 2020; von der Embse et al., 2021) and the Ofqual report are centred around the G.C.S.E. (General Certificate in Education) aged students (15-to16-year-olds) and Putwain (2020, p.6) states there is a "pressing need" for an intervention with the Year 6 age-group (10-to11-year-olds). This research contributes to closing that gap.

Test anxiety is difficult to measure (e.g., Schutz, 2014). Students who have more A trait anxiety are more likely to develop A state anxiety (Spielberger & Vagg, 1995) and this anxiety can be assessed and allowed for in access arrangements for exams. This study uses questionnaires for this age-group that can identify a student who is high in cognitive test anxiety and potentially allow for access arrangements to be made accordingly. However, there will be some students who don't have trait anxiety (Spielberger & Vagg, 1995) but who do suffer with state test anxiety and delivering an intervention in a universal fashion is one way of providing a coping strategy for students who may experience state test anxiety as well as helping students whose trait anxiety is exacerbated in evaluative contexts. The discussion for the main research questions is now concluded.

9.5 The Additional Research Questions.

Two questions arose as I studied. These are Research Questions 4 and 5. Research Question 4 asked if there was a significant difference between the groups in the intervention group where I, as a researcher delivered the intervention compared to similar groups where a class teacher delivered the intervention. Research Question 5 was called the Covid SARS COV 2 Question and asked if there was a difference in feelings of test anxiety between students who sit high stakes' paper tests or online tests?

9.5.1 Interpretation of results

The results in Research Question 4 showed that there was not a significant difference in the reduction of test anxiety to students between the interventions which I led as a teacher / researcher and the interventions that other teachers led. This result in testing the fidelity of the intervention suggests that the intervention delivered by me did not alter significantly to the intervention when delivered by other teachers. The samples for comparison were selected carefully to mirror my groups. This result supports Yang, Luo et al. (2021) who suggest that consistency of instructor between control and intervention groups did not significantly affect the effect of the testing effect.

Research Question 5 arose because of the pandemic that meant that the last part of the data collection for the control group took place online. The pre and post-test questionnaires were administered before and after a high stakes' online Math test. The results of this question showed that there was not a significant difference in the reduction of test anxiety between the students who took tests online and students who took tests on paper.

As the pandemic closed schools down in March 2020, I was waiting for one last set of pre-test and post-test Children's Test Anxiety Scale and Self-efficacy in Test Taking questionnaires that were being administered pre and post SATS (Standard Attainment Tests). I wasn't sure what to do when schools closed but, as a headteacher, I was busy moving a school to an online provision, so my priorities were with that process.

Fortunately, a teacher who had read my article (Barsham & Ellefson, 2020) got in touch with me as she was very interested in the intervention. She agreed to complete the questionnaires with her Year 6 students pre-test and post-test a high stakes' Math test that the students were taking online. The online exam was taking the place of the SATS (Standard Attainment Tests) exam.

9.5.2 Limitations

The nature of real-world research conducted in classrooms means that not all variables can be controlled for. The exact timing of the post-tests will differ depending upon when students finished sitting high stakes' tests and any day to day, school to school, child to child variables can come into play (see Chapter 10).

A possible limitation for Research Question 5 is that the students may not have considered the online test as higher stakes as a real SATs (Standard Assessment tests) exam. There may be a difference in taking a high stakes' test from a computer at home in

comparison to sitting in an exam hall and this may have affected their response values in the questionnaires. I did have the opportunity to ask the students if they had treated the online tests as seriously as a SATs test. Out of 28 responses, 15 students said ‘maybe’, six students said ‘no’ and seven students said ‘yes’.

9.5.3 Implications

The implications of the results from Research Question 4 suggest that the intervention can be delivered by any teacher. Chapter 10 includes a reflection about the intervention in this study.

For Research Question 5, one possible implication is that the student responses about test anxiety and self-efficacy in test taking were influenced by the recent cancellation of high stakes’ tests and replacement with mock, online, at home, high stakes tests. However, the independent *t* test suggests that this group’s scores did not differ significantly to a similar group that were taking actual high stakes’ tests.

9.6 Conclusion

The results suggest that the intervention may increase self-efficacy in test taking and reduce the worry element of test anxiety for 10–11-year-olds especially for students who are high in test anxiety. However, because self-efficacy increased for both the intervention and control group and worry reduced for the intervention and the control group, the results suggest that students in primary classrooms may generally be feeling more confident about taking tests. However, when measured with the Little Self-efficacy in Test Taking subscale, self-efficacy in test taking increased significantly for the intervention group and reduced for the control group. This result may be attributed to how students, especially students who are high in test anxiety respond to the negative wording of this subscale.

This study has used three theories: cognitive theory of the testing effect, test anxiety theory and self-efficacy theory to see if teaching students about the testing effect can reduce the worry or cognitive element of test anxiety. Direct instruction in the testing effect for this purpose has not been used before but empirical evidence suggests that retrieval practice can reduce test anxiety for this age-group (Agarwal et al., 2014)

The pre-test and post-test questionnaires probably enhance and hinder this experiment. It is difficult to know if the questionnaires have been completed accurately and I couldn’t control for the timing of post-test responses and they may not represent true values

depending upon when they were taken and the individuals. However, questionnaires are easy to administer and may have been an appropriate measure.

In sum and on a positive note, worry reduced significantly for students who are high in test anxiety as a result of the intervention. Furthermore, the 'Letters to Bob' are testimonies of the students learning to learn and may suggest that this is a good age to intervene with teaching students how to learn. Negatively, the results do not suggest that the intervention caused a reduction in cognitive test anxiety for the intervention group only and there may be many variables that are unaccounted for.

Using retrieval practice to learn aspects of cognitive theory may provide a coping mechanism by creating self-efficacious thoughts to manage test anxiety. I do not think that the testing effect has been used in this way, to support the pastoral care of students in education before.

The main contributions to the fields of cognitive science and education as a result of my study are: Firstly, I have provided an overarching concept that uses the cognitive theory of the testing effect to teach the testing effect to increase self-efficacy in test taking to reduce test anxiety in 10-to 11-year-olds. This concept uses retrieval practice as a pastoral support mechanism for test anxiety. In addition, I have designed a new intervention that is aimed at the self-knowledge beliefs strand of the S-REF (Self-referent executive function) model of test anxiety (Zeidner & Matthews, 2005) using cognitive theory of the testing effect that can be delivered in a classroom and is described as a 'pressing need' by Putwain. (Putwain, 2020). Furthermore, I have created two new measures: The Adapted Thoughts' subscale and the Self-efficacy in Test Taking Instrument and tested these measures for measurement invariance or partial measurement invariance which held for all subscales across timepoint, group and gender for 10-to 11-year-olds. The new measures need to be cross validated on a further sample.

Putwain (2020) refers to the need to raise the awareness of policy makers, school leaders, and school staff responsible for the pastoral care and well-being of students to the possibility of using test anxiety interventions. This study contributes an intervention aimed at the primary age-group, Year 6 to help them with all of the high stakes' tests that they have to sit and measures to identify cognitive test anxiety and self-efficacy in test taking for this age-group (Barsham & Ellefson, 2020).

10 Reflections and Conclusions

Ellefson et al. (2019) discuss the lessons that they have learned as psychologists in working with educators in schools. They mention the way that teachers work in trying something out and modifying it as they teach as being different to how psychologists conduct their experiments. The difference between good teaching and good science. As an educator and researcher, I have tried to achieve a similar type of difficult balance in this study. I have tried to design an intervention to allow for a flexible delivery (it was simple worksheets and discussion based over six weeks) and a robust measure to gauge the impact.

As an EdD student who is studying cognitive science and who has a long standing career in education, I have a foot in both of these camps and take this closing chapter as an opportunity to think out loud, to explore ideas and make recommendations to educators and to policy makers and for reform, in a way that does not support the adoption of faddish initiatives before the science is thoroughly examined (Ellefson et al., 2019) but as a result of the combination of my study and experience.

The caveat is that further studies are needed before final conclusions can be drawn and the new measures should be cross validated on another sample but some ideas for future researchers and educators and policy makers are presented here that are dedicated to improving outcomes for students who are anxious about exams.

10.1 Reflections About the Intervention

In Chapter 3, I refer to aspects of design-based experiments that I hoped to fulfil in my study of which the intervention is a key component of the design. I repeat them here and remind the reader that they are not exhaustive of ideas about design-based experiments, but they provide a good starting point for the evaluation of the intervention in my study.

Table 10.1

Evaluation of the Design of the Study / Intervention

Design-based Criteria	This Study
• Research was situated in a real educational context, the messy setting of primary classrooms (McKenney & Reeves, 2013)	• Yes, in primary classrooms, in different schools.
• The study includes the design and testing of an intervention to address a problem in practice, a real-world intervention (McKenney & Reeves, 2013)	• Yes, the problem is test anxiety. • The intervention is designed to increase self-efficacy in test taking and reduce the worry element of cognitive test anxiety.
• The intervention could be easily administered in any classroom by any adult (Ann Brown, 1992)	• I think so but I may be subject to confirmation bias. • I should have asked this specific question to teachers.
• There is an empirical investigation yielding theoretical understanding, theory building and practical applications that can inform the work of others (McKenney & Reeves, 2013)	• The research questions were answered empirically and add to test anxiety theory, the testing effect and self-efficacy theory. • The results add to the theory of test anxiety by using the testing effect to increase students' self-efficacy in test taking to reduce test taking worry and may add to schools' pastoral curricula. • The new measures and the intervention are practical applications that can inform the work of others.
• Cohen et al. (2017) - the pedagogical goal to be investigated.	• Can retrieval practice of the testing effect improve self-efficacy in test taking?
• Data collection uses multiple methods and is iterative and that the intervention is adapted to appeal to all stakeholders. Cobb et al. (2003)	• No, mainly quantitative data. • The pilot used teacher consultation about the intervention, and I trialled both questionnaires. • Consultation with Year 4 and Year 6 teachers about the pilot resulted in a few tweaks to Week 2 worksheets. • Feedback from teachers and students (Appendix, D) is valuable for future iterations.
• Cobb et al. (2003, p.9) refer to the "learning ecology" of a design experiment and the "crucibles for the generation and testing of theory".	• I tried to design an intervention that could be delivered easily in the complex settings of classrooms that was purposeful in terms of teaching the testing effect and with provision for discussion. I provided the resources for teachers.

Note. The design experiment in educational research criteria listed here is not exhaustive of its field.

Lendrum and Humphrey (2012) discuss the importance of studying the implementations of intervention and encourage the reader to move away from the idea of implementation failure to be critically evaluative, to ensure we learn from our mistakes. I have tried to be critically evaluative in this section.

As a teacher and then as a researcher within the post-positivist paradigm, I am very aware of the problems of delivering a classroom intervention and that is why the design of this intervention was simple: pre and post-test questionnaires, worksheets that could be handed out, the option for the pupils to complete them or just to read them ready for a discussion and importantly no requirement for assessment by teachers. I tried to minimize teacher workload by keeping the intervention discussion based but the teachers did need to read the resources and understand the content. It is my opinion that the intervention was easy to deliver, and teachers may have felt differently about the administration of the intervention.

As a teacher, I am very aware of time constraints and the changing nature of the school day. Staff absences, cover work and last-minute play rehearsals are just a few examples of the variables in a school context. Spontaneity of education is a key aspect of primary schools as learning opportunities are seized as they arise such as a snow falling or a wasp arriving in the classroom. These happenings (and many more) can affect the delivery of an intervention.

10.1.1 What Went Well

I have used a favourite, summative and formative classroom assessment technique to evaluate the intervention: what went well, even better if, and next steps.

The idea of using a semi-narrative to convey cognitive theory, featuring cognitive scientists as characters in a learning story, so that students can learn how to learn, is innovative and the majority of students understood the testing effect. (see Appendix D)

There was very little missing data on the questionnaires themselves and most parents were happy for their children to take part. One school in the control group had some parents who did not wish their children to complete the questionnaires. I provided the template of the letter to parents, to obtain informed consent, to the schools and also an accompanying flyer and the message that I was available if any parent would like further information. Schools then used their administration teams to send out the opt in permission letter and collated responses through parent emails. I was in communication with the headteacher or teacher

from the schools and they confirmed with me when they had received permissions. I was in regular communication with the schools throughout the processes by email and phone calls.

Schools were given the option of being an intervention school or a control school where just the questionnaires were completed, and this latter option suited some schools. Through conversations and regular communication, schools knew that they had the option of delivering the intervention, the following year.

In addition, I provided a comprehensive manual for the delivery of the intervention (see Appendix B). headteacher permissions were sought and gained and where I was the headteacher, I had the approval of my line manager. I offered further training, power points and overviews of the research to any school who wanted to complete the intervention at the outset as an addition to the teachers' manual.

I was contacted by a teacher who read my article in the IMPACT journal (Barsham & Ellefson, 2020) because she was interested in delivering the intervention to her students.

10.1.2 Even Better If

I should have asked the students and teachers if they liked the semi-narrative structure of the intervention. The teachers' questions ask for their opinions about the intervention, but I could have been more specific in my questions about the implementation of the intervention.

Two schools, who had committed to delivering the intervention had not been able to send out the permission letters when I checked, due to heavy workloads. This outcome meant that I had to recruit more schools the following year and my data collection was stalled for a year because it needs to be delivered at specific times of the academic year relating to high stakes' tests. In addition, there were three Self-efficacy in Test Taking questionnaires that were missing and one Children's Test Anxiety Scale questionnaire that I couldn't use.

Some teachers were experiencing intense workloads because of the preparation for exams and their responses to the teacher questions and their notes from the students are brief. Some teachers were not able to complete all six weeks of the intervention (see Appendix D)

Teacher 'buy in' to the delivery of the intervention was varied and perhaps I could have taken a more democratic approach consulting different schools about the design of the intervention (Ellefson et al., 2019). However, time was limited because I was in the post of headteacher in school. I had already committed to visiting the schools in person and / or with Zoom© calls and could not ask my employers for more time during the school day or expect volunteering schools to remain after school hours to see me. I did manage to visit the schools

or used Zoom© calls to speak to the headteachers and teachers for all the schools who took part in the intervention. I was able to pilot the intervention and received feedback from a Year 6 teacher (the class form teacher) and a Year 4 teacher who suggested making some changes to the Week 2 worksheet.

Lendrum and Humphrey (2012) refer to OFSTED (Office for Standards in Education, Children's Services and Skills, 2010) and the problem of initiative overload. If a teacher feels that this intervention is simply another initiative to complete, then the 'buy in' may not be there. It became apparent once I received the teacher feedback that two schools were delivering a teaching and learning initiative that was embedded into their Year 6 curriculum. It is possible that my intervention contributed towards a feeling of initiative overload. The question of additional intervention was one that I did not ask at the outset but should have done.

Lendrum and Humphrey (2012, p.636) suggest that most successful interventions work from a bottom-up approach and not a top down one. Most of the schools in the intervention group were situated in the bottom-up position as I worked with the teacher, or was the teacher, who was delivering the intervention. However, for two schools, I had met with the headteacher who agreed to deliver the intervention and then intervention was delegated. It is possible that the teacher delivering the intervention felt that this request was a top-down approach.

Scammacca et al. (2007) refer to researcher led interventions having significantly larger effect sizes than teacher led interventions. Research Question 4 is a comparison of two similar schools and similar sample sizes between me, as a researcher, and another teacher and the results suggest that there was no significant difference. This result may suggest that the intervention can be delivered by a teacher who is not a researcher.

In terms of the duration of the intervention, Willingham (2006, 2012) discusses how the length of an educational intervention can affect the effectiveness. This intervention was designed to be delivered over six weeks and consists of a 15-to-20-minute recall through discussion about knowledge gained. It was designed to be easily administered and the feedback suggests that the letter writing activities and video were enjoyed by the students, but the Week 2 worksheet was difficult to explain. The duration for optimal recall is suggested (e.g., Roediger & Karpicke, 2011) to be two days and upon reflection, the intervention could be shortened. I opted for weekly sessions because most timetabling in schools happens

weekly and also, I did not want to be over prescriptive. Time is precious in schools and often scarce, so a weekly option gives more flexibility. I should have included in my questions to teachers an option for commenting on the duration of the intervention.

10.1.3 Next Steps

If I were to run the intervention again, I would make a video and ask some students to talk about the testing effect as I think teachers and students would respond to hearing about the intervention from the students themselves.

I would alter the worksheet for Week 2 and differentiate the desirable difficulties. Some students struggled to understand spacing and interleaving but most understood the testing effect, so the spacing and interleaving content needs to be offered as an addition here.

To hear the pupil voice, I created the questions to students about the intervention but asked the teacher to collate notes from student responses. I did not want to ask the students to complete another paper questionnaire. I would like to ask students how they feel about this feedback opportunity and if they have any ideas of a better way to gather their important opinions after the intervention and high stakes' tests.

In addition, I would ask teachers to be specific about when they delivered the post-tests. I know, through conversations that most post-tests happened soon after the students had finished all of their tests and these timings did vary from school to school but I could improve the feedback by asking for the date of the last test and the date the post-test was given.

Test anxiety is difficult to measure because one cannot ask questions just before or during a test as it may affect performance (e.g., Schutz et al., 2014). Post-test responses may reflect debrief feelings of relief. If high stakes' test were to go online, then aspects of machine-based learning may be able to capture test anxiety responses more accurately (e.g., D'Mello et al., 2014). However, if the measure becomes too complex, it may not be as easy to administer in a school.

10.2 General Reflections About the Feedback, Results and Study

The feedback (see Appendix D) may make a non-scientific contribution about judgements of learning (Agarwal et al., 2008; Bjork et al., 2013; Bjork & Bjork, 2019). Bjork and Bjork (2019) suggest that even when students are told that blocking or massed practice is not the most optimal way of learning, the students struggle to believe that the testing effect is more effective. The majority of students involved in this study did understand the benefits of the testing effect. It may be that teaching the testing effect to a younger age-group is more

effective than teaching it to older students. As I have not measured judgements of learning, this comment is not a scientific statement, but an opinion based on the feedback from students that I taught and from the general feedback. Most of the literature about judgements of learning is with older students and mainly college aged students. However, Lipowski et al., (2013) suggest third grade students (U.S., 8-to 9- year-olds) understood the testing effect.

The two concepts of retrieval practice and judgements of learning are linked as students need to know why they are using a specific strategy and need to know how it works (e.g., Pekrun, 2006); the intervention in this study gives the 'why'. The feedback from students and teachers (see Appendix D) suggests that students understood the benefits of the testing effect and some examples of 'Letters to Bob' are, in my opinion, good examples of students' knowledge about their learning. Students who are high in control are those that can understand why their strategy resulted in success (Martin, 2002, 2007; Putwain, 2018)

The results from the control group for both data sets suggest that students may feel confident about taking tests and this result could reflect a change in classroom practice illustrated recently through the incorporation of desirable difficulties into the U.K. Initial Teacher Training Core Framework (2020) and potentially other resilience building strategies in schools such as Growth Mindset initiatives (Dweck, 2008). It is possible that these whole school initiatives do work in increasing academic buoyancy (Martin & Marsh, 2008) and resilience (Putwain, Nicholson et al., 2013) for the majority of the students but that they may not be specific enough to target debilitative test anxiety that affects between 10-20% of a school population. The results of this study suggest that a specific test anxiety intervention using the cognitive theory of the testing effect significantly reduced test anxiety for students who are high in test anxiety.

Learning how to learn and in this study, learning about the testing effect is a way of supporting students' confidence. High stakes' tests continue to be on the educational horizon and test anxiety continues to be a real problem for some students, approximately 10-20% (Putwain & Daly, 2014) and potentially more so in the aftermath of the pandemic and its effects on the mental health of students of all ages.

Putwain et al. (2020) in their recent paper about students' emotional responses to the cancelled high stakes' tests in 2020 include student comments about how much pressure they feel from parents and teachers about exams. This paper highlights the need for effective interventions and teacher and parent training in awareness of these issues.

In addition, the recent publication by Ofqual (Howard, 2020) that reviews the literature about anxiety for educational assessments suggests that the issue of test anxiety as being debilitative is being acknowledged at governmental level. However, not enough research has been done to see how using desirable difficulties and specifically the testing effect, can be used to better manage test anxiety. One exception is Agarwal et al. (2014) whose study focused on middle school and high school students and the impact of retrieval practice on test anxiety. Smith et al. (2016) discuss the positive results of retrieval practice in reducing stress, but the participants are college aged students in this study.

There is no mention of using desirable difficulties to manage test anxiety in the Ofqual publication (Howard, 2020). There is reference to using writing strategies to manage test anxiety as the publication cites Ramirez & Beilock (2011). Writing is a key aspect of the intervention in this study.

This study is innovative in its use of cognitive theory, the testing effect to increase self-efficacy in test taking, increasing confidence, and reducing worry. This study is innovative in that it is about a test anxiety intervention with the upper primary age-group and does not use retrieval practice for academic recall but for teaching students a strategy to help manage test anxiety. The intervention may need some adaptations in the light of its review in this thesis and through the viva process but the idea of teaching students about the testing effect to make them feel more self-efficacious may be a good one. It gives students the 'why' as to why they should learn to use the testing effect.

10.3 Recommendations – General, Policy, Reform?

Changes are needed to manage test anxiety. These should include core curriculum change. Changes to the U.K. Initial Teacher Training Framework (2020) to include aspects of desirable difficulties (Bjork, 1975) is an important step but teaching students, as well as training teachers, about the testing effect and desirable difficulties is important. In addition, using the testing effect and desirable difficulties not just for subject content teaching and learning but in ways that boost self-efficacy by making links to why aspects of learning are what they are, as exemplified in the intervention in my study, means that these important learning strategies can be used to support a pastoral curriculum as well as an academic one.

Secondly, the use of increased interventions that use writing about tests, especially for 10-to 11-year-olds are, in my opinion as a teacher and researcher, manageable and effective strategies for schools to adopt.

Thirdly, teacher, parent and peer training in the use of fear appeals or indeed, the opposite of fear appeals such as using reassurances like 'your testing routes are well-oiled' are needed. The fear appeals research (e.g., Putwain et al., 2016, 2016; Putwain & Best, 2011) shows that fear appeals can work in motivating some students to do well. It is about gauging their effect at an individual level and not using it as a blanket resource in a classroom because for students who are high in test anxiety, fear appeals can increase feelings of test anxiety (Putwain et al., 2016). Alternatively, changes to how students are tested could be considered and implemented so that students can control tests more and thereby reduce trait and state test anxiety (Zeidner, 1998).

10.3.1 Writing About Test Anxiety

The idea of writing down feelings and worries ahead of exams (e.g., Lang & Lang, 2010; Mckeachie, 1951; Ramirez & Beilock, 2011) may be a workable strategy to alleviate test anxiety. The letter to Bob Bjork recalling the benefits of learning about the testing effect and the task of writing to a friend to give advice about test anxiety were activities in the intervention that the students and teachers felt benefited them and enjoyed (see Appendix D). The narrative structure of the intervention was a deliberate choice as I felt this model would be a medium that students of this age would resonate with. In addition, writing to a friend was a deliberate choice as, as a classroom teacher, I know how keen students are to help and to advise their friends.

Small acts of writing could strengthen the resilience of students who have to take high stakes' tests in the near future. (e,g., Kliewer et al., 2011; Lepore & Smyth, 2002; Lepore et al., 2002). A counter theory to this one is found in Fivush et al. (2007) whose study with 9-to 13-year-olds suggests that writing about negative emotions may be detrimental to some students. However, the feedback from students and teachers in this study suggest that, in general, the writing activities were enjoyed by the students. My own experience was that the students were excited to write a letter to Bob in California "Will you really send them Mrs. Barsham?" and also the students felt enabled from the intervention to write advice to a friend.

10.3.2 Training About Fear Appeals

Putwain's call for attributional training for teachers (Putwain & Aveyard, 2018) is, in my opinion, a sensible suggestion so that teachers have the training to support students' successes and failures. Schutz et al. (2014) suggest the impact of teachers upon test anxiety is an important area for investigation. The role of a teacher in shaping attribution is explored in Hattie and Timperley (2007). Easily accessible training programmes for teachers should be available so that an awareness created about what fear appeals can do to students (and how to gauge their use) and the opposite, such as how to use test anxiety interventions and the cognitive theory of the testing effect to improve self-efficacy in test taking.

Given the recent loss of education for many students due to the pandemic, the use of fear appeals may become more intense as teachers struggle to get classes across the assessment line. For year groups that have potentially missed huge chunks of curriculum with high stakes' tests hovering on the horizon, the pressure and stress may be evident for all. The intervention in this study can be delivered by a classroom teacher and could help students, especially students who are high in test anxiety to believe in their test taking skills.

Test anxiety interventions may improve student well-being (e.g., Weems et al., 2015; Putwain et al., 2021; Steinmeyr et al., 2016) and there is a "pressing need" (Putwain, 2020) for a test anxiety intervention for students in Year 6. There should be an increased use of test anxiety interventions and an increase in training for teachers so that they can understand and manage students' test anxiety effectively.

10.3.3 Proposal for Curricula Reform

Retrieval practice seems to facilitate learning for students who have low working memory capacity (Agarwal & Finley, 2017) and test anxiety (Agarwal et al., 2014). It is now considered such an important aspect of learning science that it is included in the Initial Teacher Training Core Content Framework (2020).

Using retrieval practice in a pastoral capacity to give a strategy for managing test anxiety has not been done before and could be included onto P.H.S.E. (Personal, Health, Social and Economic) or other pastoral curricula. In addition, an academic curriculum or an extra-curricular programme that is designed to teach students how to learn and how to judge their learning accurately is likely to improve student outcomes.

I suggest in 2.1.3 that teaching students, key aspects about how learning works is integral to developing learning and improving outcomes for students. This suggestion came

from the research that discusses judgements of learning (e.g., Bjork et al., 2013) and students' inability to judge what they know. I suggest teaching students about memory and biases such as foresight bias and hindsight bias. In my opinion, if these learning how to learn (Bjork, 2014) techniques were embedded in curricula then this provision would reduce the need for test anxiety and maybe other types of intervention. The metaphor that springs to mind is of treating the root cause rather than the symptoms. What better education could we give than teaching students how to learn?

There is a need to develop training for teachers (and parents) and curricula for students that build upon the recent additions of desirable difficulties to the U.K. Initial Teacher Training framework (2020) and teach more about how memory works and the different learning biases of foresight and hindsight bias. There is a need to create an age relative, year upon year learning skills incremental curriculum using desirable difficulties. The following is a simple suggestion and would need much more refinement but provides a basic idea: Year 5/6, the testing effect, and an introduction to some Desirable Difficulties; Year 7-9, the testing effect, interleaving, spacing, moving rooms and elaboration; Year 10, the testing effect, retrieval fluency and biases; Year 11, the testing effect, retrieval fluency and biases; Year 12, Desirable Difficulties, New Theory of Disuse, Retrieval Induced Forgetting and Year 13 as per Year 12. This idea could be continued and added to in Higher Education too.

The recent paper by Ofqual (Howard, 2020) is an overview of the effects of test anxiety. Whilst 'academic skill building' is mentioned as an intervention for test anxiety, the use of desirable difficulties (to treat the root of the problems rather than the symptoms) are not mentioned and is potentially a valuable missing link.

This study is unique in that it uses the desirable difficulty of the testing effect to create test taking self-efficacy to explain the academic benefits of retrieval practice and thereby create a pastoral support system. I am suggesting that retrieval practice should not be limited to subject content but should incorporate rehearsal, retrieval and recall of confidence building, learning strategies that can be easily recalled to create self-efficacy in a given situation. 'Your testing routes are well-oiled'.

In sum, I suggest a curriculum or extra- curriculum that teaches desirable difficulties and a pastoral curriculum that uses the testing effect to manage test taking anxiety and potentially contributes towards student well-being as well.

10.4 Future Directions

Because of the scarce amount of literature about test anxiety and retrieval practice and the almost non-existent (Barsham & Ellefson, 2020) literature for students who are high in test anxiety for 10-to 11-year-olds, the ideas for future studies are extensive and take the form of a list here: Future studies should cross-validate the new measures from this study with other samples of Year 6 students; future studies could explore the link between self-efficacy in test taking and well-being for this age-group; future studies should examine if upper primary is an optimal age for learning about the testing effect in comparison to younger or older year groups and if 10-to11-year-olds can make accurate judgements of learning; future studies should examine the effect of writing about worries and / or positive written messages on social media on test anxiety for 10-to 11-year-olds; future studies should include a longitudinal aspect to see for how long students remember "study, test, test, test" (Roediger & Karpicke, 2006b) as a test taking strategy; future studies should acknowledge the changing landscapes in education and in children and explore thoroughly the nature of the test anxiety construct at the time of measuring it (and create appropriate measures) and explore further the possible differences of test anxiety and paper and online tests; further studies should investigate the outcomes for pupils of attributional training of teachers and reductions in fear appeals and the replacing of negative language with positive 'your testing routes are well oiled'. Furthermore, research into how students who are high in test anxiety respond to wordings on questionnaires would be useful.

In addition, Imundo et al. (2020) discuss research about the effectiveness of recall of knowledge in different rooms. Future studies should examine if students, who have the opportunity to revise / recall information for exams in the exam room itself are less anxious. It may be that this process would act as a type of desensitisation programme. I would very much like to carry out this research as I regularly encourage my students to practise recalling their knowledge in the exam room.

Finally, because test anxiety is a very difficult construct to measure and relies heavily on self-report questionnaires, future studies should consider different approaches to measuring the test anxiety construct such as complementary, qualitative data collection or within person analyses (Murayama et al., 2017)

10.5 Conclusion About the Intervention and Final Reflections

It is difficult to compare this study to another because there is not one like it. However, the intervention may have provided a type of exercise that does not teach academic content but does target a student's belief and may have an effect on educational achievement. (Yeager & Walton, 2011). In the case of my study, achievement would be increased feelings about self-efficacy in test taking because of knowing how the testing effect works. My intervention may be similar to the type of social psychological intervention in Blackwell, Trzesniewski and Dweck (2007) who found that middle school students that attended an eight-session workshop that taught them that the brain is like a muscle and grows with effort benefited from a sharp increase in math achievement for the rest of the school year. Yeager and Walton (2011, p. 293) caution the reader about attributing magical qualities to these social psychological interventions and suggest they are dependent upon "capacities, meaning and recursive processes" but they do recognize that long lasting gains are achieved, and psychological interventions have a demonstrated potential to address fundamental problems.

By combining aspects of the three fields of cognitive science, test anxiety and self-efficacy, I designed an intervention to increase feelings of self-efficacy in test taking to alleviate test anxiety in 10-to11-year-olds. The intervention may be the missing piece of the jigsaw puzzle (feedback to me from a colleague) because explaining why the testing effect works, so that students can make the link to why they are being tested (recall aids future recall) in turn increases self-efficacy in test taking.

The intervention can be delivered by a classroom teacher and might be seamlessly integrated into school routines (Weems et al., 2015). The intervention may contribute to Weems et al. (2015) call for interventions that help students improve testing success and improve self-efficacy. Moreover, the intervention in this study may fulfil the "pressing need" for a Year 6 intervention for test anxiety (Putwain, 2020, p.6).

Test anxiety interventions are needed to manage test anxiety in schools and especially for the 10-20% of students who are categorized as high in test anxiety and for Year 6 (e.g., Putwain, 2020) However, test anxiety interventions need to be used in conjunction with bigger changes in education. Pastoral curricula should include learning about the testing effect for students who must sit high stakes' exams to give them a strategy for managing these tests. As a headteacher and researcher, I will encourage my students to practise study,

test, test, test and ensure that they can understand why this process benefits their learning and test taking skills.

Bjork & Bjork (2020) conclude a recent review of desirable difficulties in theory and practice by suggesting that their estimation of the potential to improve teaching and self-regulated learning by using desirable difficulties in the classroom has increased but that there are challenges. In the U.K. the inclusion of retrieval practice into the Initial Teacher Training Core Framework (2020) is a step in the right direction but teaching about desirable difficulties should be explicit and extended to the students and not simply rest with teacher pedagogy.

Finally, my view as a headteacher and researcher, is that the way forward in education is in teaching students about how memory and learning biases work, to give students the tools for learning for life. For so many years, schools have been defined by their strands of pastoral, academic and extracurricular provision. Staffing structures, websites, the school day and curricula are organised by these three core elements of education. I suggest it is time for an overarching ‘meta-curriculum’ in the form of teaching students how they learn to learn by teaching them about desirable difficulties to be adopted by schools to prepare students fully for the future.

References

American Psychiatric Association (2000) *Diagnostic and statistical manual of mental disorders* (4th ed., text rev). Author.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Author.

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger, H. L., 3rd (2017). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory 25*(6), 764–771. https://doi.org/10.1080/09658211.2016.1220579

Agarwal, P. K., D'Antonio, L., Roediger, H. L., McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition, 3*(3), 131–139. https://doi.org/10.1016/j.jarmac.2014.07.002

Agarwal, P. K., Roediger, H. L., McDaniel, M. A., & McDermott, K. B. (2010). Improving Student Learning through the Use of Classroom Quizzes: Three Years of Evidence from the Columbia Middle School Project. (ED513945) *Society for Research on Educational Effectiveness*. ERIC. https://eric.ed.gov/?id=ED513945

Agarwal, P.K., Karpicke, J.D., Kang, S.H., Roediger, H., & McDermott, K. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*, 861-876. https://doi.org/10.1002/acp.1391

Agarwal, P.K., & Bain, P.M. (2019) *Powerful teaching*. Jossey Bass, A Wiley Brand.

Albin, R. W., Lucyshyn, J. M., Horner, R. H., & Flannery, K. B. (1996). Effects of Behaviour Support Team Composition for Technical Adequacy and Contextual Fit for Behavior Support Plans. In L. K. Koegel, R. L. Koegel, G. Dunlap (Eds.), *Positive behavioral support: Including people with difficult behavior in the community* (pp. 81- 98). Brookes.

Alpert, R., & Haber, R.N. (1960). Anxiety in academic achievement situations. *Journal of abnormal and social psychology. Vol 61*(2), 207-215 https://doi.org/10.1037/h0045464

Anderson, M.C., Bjork, R.A., & Bjork, E.L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, Vol 20*(5) 1063-1087 https://doi.org/10.1037//0278-7393.20.5.1063

Anderson, T. & Shattuck, J. (2012) A decade of progress in education research. *Educational Researcher, Vol 41*(1) pp16-25.. Sage Publications. https://doi.org/10.3102/0013189X11428813

Anderson, B. S., Wennberg, K., & McMullen, J. S. (2019). Enhancing quantitative theory-testing entrepreneurship research. *Journal of Business Venturing*, *34*(5), 105928. https://ideas.repec.org/a/eee/jbvent/v34y2019i55.html

Antonakis, J., ; Bastardoz, N., & Rönkkö, M. (2021). On ignoring the random effects assumption in multilevel models : review, critique, and recommendations. *Organizational Research Methods, 24*(2), 443-483. https://doi.org/10.1177%2F1094428119877457

Aslan, A., & Bäuml, K. T. (2016). Testing enhances subsequent learning in older but not in younger elementary school children. *Developmental science*, *19*(6), 992–998. https://doi.org/10.1111/desc.12340

Aydin, U., & Bulgan, G. (2017). Çocuklarda Sınav Kaygısı Ölçeği'nin Türkce uyarlaması [Adaptation of Children's Test Anxiety Scale to Turkish]. *İlköğretim Online [Elementary Education Online]*, *16*(2), 887-899. http://dx.doi.org/10.17051/ilkonline.2017.304742

Baker, S., Ellefson., M.E., Tsapali, M., Barsham, H. (2020) Cambridge Primer for P.G.C.E. Students Teacher Training Resource, unpublished.

Bandura, A. (2006) Guide for constructing self-efficacy scales in F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents Vol. 5*, pp. 307-337. Information Age Publishing.

Bandura, A. (1997). *Self-efficacy. The exercise of control.* Stanford University: W.H.Freeman/ Times Books / Henry Holt & Co

Bandura, A. (1986a) *Social foundations of thought and action: A social cognitive theory*. Prentice -Hall.

Bandura, A. (1989) Human agency in social cognitive theory. *American Psychologist, 44,* 1175-1184.https://doi.org/10.1037/0003-066X.44.9.1175

Bandura, A. (1986b) The explanatory and predictive scope of self-efficacy theory. *Journal of Clinical Psychology 4*, 359-373. https://doi.org/10.1521/jscp.1986.4.3.359

Bandura, A. (1978). The self-system in reciprocal determinism. *American Psychologist, 33* (4), 344–358. https://doi.org/10.1037/0003-066X.33.4.344

Bandura, A. (1977). *Social learning theory.* Prentice-Hall.

Bartlett, M. S. (1951). The effect of standardization on a Chi-square approximation in factor analysis. *Biometrika, 38*(3/4), 337-344. https://doi.org/10.1093/biomet/38.3-4.337

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences, 13*(1), 1–14.https://doi.org/10.1207/s15327809jls1301_1

Barsham, H., & Ellefson, M.R. (2020) Can teaching upper primary about the testing effect increase feelings of confidence about test taking? *Journal of the Chartered College of Teachers, Impact 8.* https://impact.chartered.college/article/can-teaching-upper-primary-testing-effect-increase-feelings-confidence-test-taking/

Benson, J., & Hocevar, D. (1985) The impact of item phrasing on the validity of attitude scales for elementary children. *Journal of Educational Measurement 22*(3) 231-240 https://doi.org/10.1111/j.1745-3984.1985.tb01061.x

Bialosiewicz, S., Murphy, K., & Berry, T. (2013) *An Introduction to Measurement Invariance Testing: Resource Packet for Participants Do our Measures Measure up? The Critical Role of Measurement Invariance Demonstration Session*, American Evaluation Association, Washington, DC Claremont Evaluation Center - School of Social Science, Policy, & Evaluation Claremont Graduate University. http://comm.eval.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=63758fed-a490-43f2-8862-2de0217a08b8

Birenbaum, M., & Nasser, F. (1994) On the relationship between test anxiety and test performance. *Measurement and Evaluation in Counseling and Development*, *v27* n1 p.293-301. (EJ487656) ERIC. https://eric.ed.gov/?id=EJ487656

Blackwell, L. S., Trześniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: a longitudinal study and an intervention. *Child development*, *78*(1), 246–263. https://doi.org/10.1111/j.1467-8624.2007.00995.x

Bjork, R.A. & Bjork, E.L. (2020) Desirable difficulties in theory and practice, *Journal of Applied Research in Memory and Cognition, 9*(4), 475-479 UCLA Bjork Learning and Forgetting Lab. https://bjorklab.psych.ucla.edu/wp-content/uploads/sites/13/2021/01/RABjorkELBjorkJARMAC2020ForPostingSingleSpaced.pdf

Bjork, R. A., & Bjork, E. L. (2019). The myth that blocking one's study or practice by topic or skill enhances learning. In C. Barton (Ed.), *Education Myths: An Evidence-Informed Guide for Teachers.* John Catt Educational Ltd. https://bjorklab.psych.ucla.edu/wp-content/uploads/sites/13/2020/01/BjorkBjorkEducatinMythChapterPublishedFormSept2019.pdf

Bjork, R. A., & Bjork, E. L. (2019). Forgetting as the friend of learning: Implications for teaching and self-regulated learning. *Advances in Physiology Education*, *43*(2), 164-167. https://doi.org/10.1152/advan.00001.2019

Bjork, R. A. (2015). *Forgetting as a friend of learning*. In D. S. Lindsay (Ed.) & C. M. Kelley (Trans.) & A. P. Yonelinas, H. L. Roediger II (Eds.), *Psychology Press festschrift series. Remembering: Attributions, processes, and control in human memory: Essays in honor of Larry Jacoby* (p. 15–28). Psychology Press. https://bjorklab.psych.ucla.edu/wp-content/uploads/sites/13/2016/11/RABjork_JacobyFestschriftChapterFigsEmbedded052014.pdf

Bjork, R. A., & Kroll, J. F. (2015). Desirable Difficulties in Vocabulary Learning. *The American journal of psychology*, *128*(2), 241–252. https://doi.org/10.5406/amerjpsyc.128.2.0241

Bjork, R.A. (2014, March 30) Learning to Learn. Conversation with Bob Bjork (Part 1) You Tube [video] https://www.youtube.com/watch?v=fTtbp6TyBrI

Bjork, R.A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: beliefs, techniques and illusions. *Annual Review of Psychology, 64,* 417-44 https://doi.org/10.1146/annurev-psych-113011-143823

Bjork, R.A. (2012, July 13) Robert Bjork- using our memory shapes our memory [video]YouTube https://www.youtube.com/watch?v=69VPjsgm-E0&t=4s

Bjork, E.L., & Bjork, R.A. (2011). *Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning*. In M.A. Gernsbacher, R.W. Pew, & J.R. Pomerantz (Eds.), & FABBS Foundation, *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp.56-64). Worth Publishers. https://bjorklab.psych.ucla.edu/wp-content/uploads/sites/13/2016/04/EBjork_RBjork_2011.pdf

Bjork, R. A. (2011). On the symbiosis of remembering and forgetting and learning. In A.S. Benjamin (Ed.), Successful remembering and successful forgetting. A festschrift in honor of Robert A. Bjork. (pp1-17). Psychology Press, Taylor & Francis Group. https://bjorklab.psych.ucla.edu/wp-content/uploads/sites/13/2016/07/RBjork_2011.pdf

Bjork, R. A. (1994a). *Memory and metamemory considerations in the training of human beings.* In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (p. 185–205). The MIT Press. https://bjorklab.psych.ucla.edu/wp-content/uploads/sites/13/2016/07/RBjork_1994a.pdf

Bjork, R. A., & Bjork, E. L. (1992). *A new theory of disuse and an old theory of stimulus fluctuation*. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of William K. Estes, Vol. 1. From learning theory to connectionist theory; Vol. 2. From learning processes to cognitive processes* (p. 35–67). Lawrence Erlbaum Associates, Inc. https://bjorklab.psych.ucla.edu/wp-content/uploads/sites/13/2016/07/RBjork_EBjork_1992.pdf

Bjork, R.A. (1975). *Retrieval as a memory modifier: An interpretation of negative recency and related phenomena*. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123-144). Lawrence Erlbaum pp.123-144. https://bjorklab.psych.ucla.edu/wp-content/uploads/sites/13/2016/07/RBjork_1975.pdf

Blaikie, N. (2010). *Designing social research.* Polity Press

Bögels, S. M., Alden, L., Beidel, D. C., Clark, L. A., Pine, D. S., Stein, M. B., & Voncken, M. (2010). Social anxiety disorder: questions and answers for the DSM-V. *Depression and anxiety*, *27*(2), 168–189. https://doi.org/10.1002/da.20670

Bolger, N. (1990). Coping as a personality process: A prospective study. *Journal of Personality and Social Psychology, 59*(3), 525–537. https://doi.org/10.1037/0022-3514.59.3.525

Boone, H.N., & Boone, D. (2012). Analyzing Likert Data. *The Journal of Extension, 50*. Eric.ed.gov. https://eric.ed.gov/?id=EJ1042448

Bracht, G. H., & Glass, G. V. (1968). The External Validity of Experiments. *American Educational Research Journal, 5*(4), 437–474. https://doi.org/10.3102/00028312005004437

Bradley, B. A., & Reinking, D. (2011). A formative experiment to enhance teacher-child language interactions in a preschool classroom. *Journal of Early Childhood Literacy, 11*(3), 362–401. https://doi.org/10.1177/1468798411410802

British Educational Research Association (2018) *Ethical guidelines for educational research*. 4th edn. https://www.bera.ac.uk/researchers-resources/publications/ethical-guidelines-for-educational-research-2018

Brown, A. (1992) Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings*, Journal of the Learning Sciences, 2:2,*141-178. https://doi.org/10.1207/s15327809jls0202_2

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, *21*(2), 230 258. https://doi.org/10.1177/0049124192021002005

Brown, P.C., Roediger, H.L., McDaniel, M.A. (2014) *Make it stick, the science of successful learning.* The Belknap Press of Harvard University Press. https://doi.org/10.1080/00220671.2015.1053373

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Browne, W.J. (n.d.) *Sample Size for Multilevel Models* University of Bristol Centre for Multilevel Linear Modelling. http://www.bristol.ac.uk/cmm/learning/multilevel-models/samples.html

Buchanan, E. (2016, June 3) *R Confirmatory Factor Analysis Lecture*, Statistics of Doom. [video]. YouTube. https://www.youtube.com/watch?v=YNTr-jwOUNk&list=PLw93TUuxrFAZV3ZgsNfMFszE8OA744NnQ&index=3

Buchanan, E. (2019, September 19) *R Multigroup CFA with Lavaan Example*, Statistics of Doom. [video]. YouTube. https://www.youtube.com/watch?v=t6uHROGalBo

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(4), 918–928. https://doi.org/10.1037/0278-7393.34.4.918

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*(4-5), 514–527. https://doi.org/10.1080/09541440701326097

Brunton-Smith, I. (2019, February 18) *Random Intercept model*. [Video]. YouTube. https://www.youtube.com/watch?v=KmbtZNjvsNM

Carifio, J., & Perla, R. (2007). Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes. *Journal of Social Sciences, 3*, 106-116. https://doi.org/10.3844/jssp.2007.106.116

Carpenter, S. K. (2012). Testing Enhances the Transfer of Learning. *Current Directions in Psychological Science*, *21*(5), 279–283. https://doi.org/10.1177/0963721412452728

Carpenter S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of experimental psychology. Learning, memory, and cognition*, *35*(6), 1563–1569. https://doi.org/10.1037/a0017021

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*(2), 268–276. https://doi.org/10.3758/BF03193405

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation science: IS*, *2*, 40. https://doi.org/10.1186/1748-5908-2-40

Carver, C.S., & Scheier, M.F. (1981). *Attention and self-regulation. A control theory approach to human behaviour.* Springer-Verlag. https://doi.org/10.1007/978-1-4612-5887-2

Cassady, J.C., & Johnson, R.E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology 27*(2), 270-295. https://psycnet.apa.org/doi/10.1006/ceps.2001.1094

Cassady, J. C. (2004). The Impact of Cognitive Test Anxiety on Text Comprehension and Recall in the Absence of External Evaluative Pressure. *Applied Cognitive Psychology, 18*(3), 311–325. https://psycnet.apa.org/doi/10.1002/acp.968

Cattell R. B. (1966). The Scree Test for The Number Of Factors. *Multivariate behavioral research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Chambers, C. (2017) *The seven deadly sins of psychology* Princeton University Press

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*(4), 553–571. https://doi.org/10.1037/0096-3445.135.4.553

Chang, C., Gardiner, J., Houang, R., & Yan-Liang, Y. (2020) Comparing multiple statistical software for multiple-indicator, multiple-cause modeling: an application of gender disparity in adult cognitive functioning using MIDUS II dataset. *BMC Med Res Methodol 20* (1) 275 https://doi.org/10.1186/s12874-020-01150-4

Chapell, M. S., Blanding, Z. B., Silverstein, M. E., Takahashi, M., Newman, B., Gubi, A., & McCann, N. (2005). Test Anxiety and Academic Performance in Undergraduate and Graduate Students. *Journal of Educational Psychology, 97*(2), 268–274. https://doi.org/10.1037/0022-0663.97.2.268

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chyung, S.Y., Barkin, J. R., and Shamsy, J. A. (2018). "Evidence-Based Survey Design: The Use of Negatively Worded Items in Surveys". *Performance Improvement, 57*(3), 16-25. http://dx.doi.org/10.1002/pfi.21749

Cizek, G.J., & Borg, S.S. (2010). *Addressing test anxiety in a high stakes' environment.* Corwin Press.

Csikszentmihalyi, M (2004, February) *Flow the secret to happiness* [Video] TED https://www.ted.com/talks/mihaly_csikszentmihalyi_flow_the_secret_to_happiness?language=en

Clark, D. A., Crandall, J. R., & Robinson, D. H. (2018). Incentives and test anxiety may moderate the effect of retrieval on learning. *Learning and Individual Differences, 63,* 70–77. https://doi.org/10.1016/j.lindif.2018.03.001

Cobb, P., Confrey, J., DiSessa, A., Lehrer, R., & Schauble, L. (2003). Design Experiments in Educational Research. *Educational Researcher, 32*(1), 9-13. https://doi.org/10.3102%2F0013189X032001009

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cohen, L., Manion, L., & Morrison, K. (2017). *Research Methods in Education* (8th ed.). Routledge. https://doi.org/10.4324/9781315456539

Costello, A., & Osborne, J.W. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation, 10* (7), 1-9. https://doi.org/10.7275/JYJ1-4868

Covington, M. V. (1984). The self-worth theory of achievement motivation: Findings and implications. *The Elementary School Journal, 85*(1), 5-20. https://doi.org/10.1086/461388

Covington, M.V. (1992). *Making the grade. A self-worth perspective on motivation and school reform.* Cambridge University Press. https://doi.org/10.1017/CBO9781139173582

Cronbach, L. J. (1951). Coefficient *alpha* and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/bf02310555

Crowson, M. (2019, July 27) Multilevel modeling using SPSS [video]. YouTube. https://www.youtube.com/watch?v=RU1ps6jaheI

Crowson, M. (2018, January 30) Two level multilevel model using SPSS (chapter 3 v1). [video]YouTube. https://www.youtube.com/watch?v=a4P_Bz0vJhE

Crowson, M (2018, January 30) Two-level multilevel model using SPSS (chapter 3 v2); HLM with random intercept plus fixed slope. [video]. YouTube. https://www.youtube.com/watch?v=iVjh2jlehZ8

Detrich, R. (1999). Increasing treatment fidelity by matching interventions to contextual variables within the educational setting. *School Psychology Review, 28*(4), 608–620. https://doi.org/10.1080/02796015.1999.12085988

Diemer, M.A. (2019, November 14) Why don't people use MIMIC models more often. Blalock lecture series. [video].YouTube. https://www.youtube.com/watch?v=cFoaJgeHQZY

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology, 105*(4), 1082–1099. https://doi.org/10.1037/a0032674

D'Mello, S., Lehman, B., Graesser, A., Pekrun, R. (2014) Confusion can be beneficial for learning. Learning and Instruction vol.29, 153-170. https://doi.org/10.1016/j.learninstruc.2012.05.003

DeWinstanley, P. A., & Bjork, E. L. (2004). Processing strategies and the generation effect: implications for making a better reader. *Memory & cognition*, *32*(6), 945–955. https://doi.org/10.3758/bf03196872

Dirkx, K.J.H., Camp, G., Kester, L., Kirschner, P.A. (2019) Do secondary students make use of effective study strategies when they study on their own? *Applied Cognitive Psychology 33* (5) 952-957. https://doi.org/10.1002/acp.3584

Donolato, E., Marci, T., Altoè, G., & Mammarella, I. C. (2020). Measuring test anxiety in primary and middle school children: Psychometric evaluation of the Test Anxiety Questionnaire for Children (TAQ-C). *European Journal of Psychological Assessment, 36*(5), 839–851. https://doi.org/10.1027/1015-5759/a000556

Dugard, P., & Todman, J. (1995). Analysis of pre-test-post-test control group designs in educational research. *Educational Psychology, 15*(2), 181–198. https://doi.org/10.1080/0144341950150207

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving Students' Learning with Effective Learning Techniques: Promising Directions from Cognitive and Educational Psychology. *Psychological science in the public interest: a journal of the American Psychological Society*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. *Health education research 18*(2), 237–256. https://doi.org/10.1093/her/18.2.237

Dweck, C. S. (2008). *Mindset: the new psychology of success*. Ballantine Books

Ebbinghaus, H (1885). *Memory: A contribution to experimental psychology* (H.A. Ruger & C.E. Bussenius, Trans.) (Original work published 1885).

Eklöf, H., & Nyroos, M. (2013). Pupil perceptions of national tests in science: Perceived importance, invested effort, and test anxiety. *European Journal of Psychology of Education, 28*(2), 497–510. https://doi.org/10.1007/s10212-012-0125-6

Ellefson, M.R., Baker, S.T., Gibson, J.L. (2019) Lessons for successful cognitive developmental science in educational settings: The case of executive functions. *Journal of Cognition and Development 20:2,* 253-277. https://doi.org/10.1080/15248372.2018.1551219

Epskamp, S., Stuber, S., Nak, J., Veenman, M., Jorgensen., T.D. (2019) Semplot in RStudio© . https://github.com/SachaEpskamp/semPlot

Ergene, T. (2003). Effective interventions on test anxiety reduction. A meta-analysis. School *Psychology International, Vol. 24,* 313-328. https://doi.org/10.1177/01430343030243004

Estes, W. K. (1955b). Statistical theory of distributional phenomena in learning. *Psychological Review,62,* 369-377. https://doi.org/10.1037/h0046888

Eysenck, M.W., Derakshan, N., Santos, R., & Calvo, M.G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion. Vol 7.* 336-353. https://doi.org/10.1037/1528-3542.7.2.336

Eysenck, M.W., & Calvo, M.G. (1992). Anxiety and Performance: The Processing Efficiency Theory. *Cognition and Emotion 6*(6) 409-434. https://doi.org/10.1080/02699939208409696

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Field., A. (2012) *Discovering statistics using IBM SPSS Statistics* (4th ed.) Sage

Field., A., Miles, J., & Field, Z. (2012) *Discovering statistics using R.* Sage

Field., A. (2009) *Discovering statistics using SPSS* (3rd ed.) Sage

Field, A. & Hole, G. (2003) *How to design and report experiments*. Sage

Finch, W.H., & French, B.F. (2015) Latent variable modeling with R (1st ed.) Routledge

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1*(3), 288–299. https://doi.org/10.1037/0096-1523.1.3.288

Fivush, R., Marin, K., Crawford, M., Reynolds, M., & Brewin, C. R. (2007). Children's narratives and well-being. *Cognition and Emotion, 21*(7), 1414–1434. https://doi.org/10.1080/02699930601109531

Flay B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive medicine*, *15*(5), 451–474. https://doi.org/10.1016/0091-7435(86)90024-1

Fraser, M.W., Richman, J.M., Galinsky, M.J., & Day, S.H. (2009). *Pocket guides to social work research. Intervention research: Developing social programmes.* Oxford University Press. https://psycnet.apa.org/record/2009-04485-000

Galassi, J. P., Frierson, H. T., & Sharer, R. (1981). Behavior of high, moderate, and low test anxious students during an actual test situation. *Journal of Consulting and Clinical Psychology, 49*(1), 51–62. https://doi.org/10.1037/0022-006X.49.1.51

Garland, R. (1991). The Mid-Point on a Rating Scale: Is it Desirable? *Marketing Bulletin* 66-70 https://www.semanticscholar.org/paper/The-Mid-Point-on-a-Rating-Scale%3A-Is-it-Desirable-Garland/e9eb4f1725b0d3923e7e1ade8efda2aabb739c49

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models: Statistical graphics for research and presentation*. p551-564 Cambridge University Press. https://doi.org/:10.1017/CBO9780511790942.033

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling, 3*(1), 62–72. https://doi.org/10.1080/10705519609540030

Goonan, B. (2003). Overcoming Test Anxiety: Giving Students the Ability to Show What They Know. (ED480053) ERIC. https://eric.ed.gov/?id=ED480053

Goldacre, B. (2013) Building evidence into education. Gov. U.K. http://media.education.gov.uk/assets/files/pdf/b/ben%20goldacre%20paper.pdf

Gonzales, J. E., & Cunningham, C. A. (2015, August). The promise of pre-registration in psychological research. *Psychological Science Agenda*. http://www.apa.org/science/about/psa/2015/08/pre-registration

Goss-Sampson, M.A. (2019) JASP (2nd ed.). *A statistical guide to JASP*. https://static.jasp-stats.org/Statistical%20Analysis%20in%20JASP%20-%20A%20Students%20Guide%20v0.10.2.pdf

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430–450. https://doi.org/10.1037/1082-989X.6.4.430

Grix, J. (2010). *Palgrave research skills. The foundations of research*. Palgrave Macmillan.

Guttman, L. (1954). *A new approach to factor analysis: the Radex.* In P. F. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences* (p. 258–348). Free Press.

Guyon, H., Kop, J.-L., Juhel, J., & Falissard, B. (2018). Measurement, ontology, and epistemology: Psychology needs pragmatism-realism. *Theory & Psychology*, *28*(2), 149–171. https://doi.org/10.1177/0959354318761606

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika, 66*(3), 373–388. https://doi.org/10.1007/BF02294440

Harlen, W., Crick, R. D., Broadfoot, P., Daugherty, R., Gardner, J., James, M., & Stobart, G. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning. EPPI-Centre, University of London. http://hdl.handle.net/1893/19607

Hartley, J. (2014). Some thoughts on Likert-type scales. *International Journal of Clinical and Health Psychology, 14* (1) 83-86. https://doi.org/10.1016/S1697-2600(14)70040-7.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hattie, J., Yates, G. Visible Learning (2014) *Visible learning and the science of how we learn.* Routledge

Hembree, R. (1988). Correlates, causes, effects and treatment of test anxiety. *Review of educational research, Vol 58,* 47-77. https://doi.org/10.3102/00346543058001047

Heck, R. H., Thomas, S. L., & Tabata, L. N. (2014). *Quantitative methodology series. Multilevel and longitudinal modeling with IBM SPSS* (2nd ed.). Routledge/Taylor & Francis Group.

Herzer, F., Wendt, J., & Hamm, A. O. (2014). Discriminating clinical from nonclinical manifestations of test anxiety: A validation study. *Behavior Therapy, 45*(2), 222–231. https://doi.org/10.1016/j.beth.2013.11.001

Hinze, S.R., & Rapp, D.N. (2014). Retrieval (sometimes) enhances learning: Performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology, Vol 28* 597-606. https://doi.org/10.1002/acp.3032

Hobbiss, M. [@mikehobbiss]. (2021, March 14th) I think that this might be the most useful single paper [image of paper] [Tweet]. Twitter. https://twitter.com/mikehobbiss/status/1371220459149217797?s=21

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185. https://doi.org/10.1007/BF02289447

Horn, E & Ashbaugh, E.J. (1920) Lippincott's Horn-Ashbaugh speller for grades 1-8. Lippincott. https://www.gutenberg.org/files/33826/33826-h/33826-h.htm

Howard. E. (2020) *A review of the literature concerning anxiety for educational assessments*, Gov. U.K. Ofqual. https://www.gov.uk/government/publications/a-review-of-the-literature-on-anxiety-for-educational-assessments

Hu. L., & Bentler, P.M. (1999) Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Structural Equation Modeling: *A Multidisciplinary Journal, 6* (1) 1-55 https://doi.org/10.1080/10705519909540118

Hulleman, C.S. & Cordray, D.S. (2009) Moving from the Lab to the Field: The Role of Fidelity and Achieved Relative Intervention Strength, *Journal of Research on Educational Effectiveness, 2:1,* 88-110. https://doi.org/:10.1080/19345740802539325

Imundo, M.N., Pan, S.C., Bjork, E.L., Bjork, R.A. (2020) Where and how to learn: The interactive benefits of contextual variation, restudying, and retrieval practice for learning. *Quarterly Journal of Experimental Psychology* https://doi.org/10.1177/1747021820968483

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired associate learning. *Journal of Experimental Psychology Vol.83*(2, Pt.1) 340-344. https://doi.org/10.1037/h0028541

ITT Core Content Framework (Initial Teacher Training Framework) Educational Endowment Fund. Department For Education https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/974307/ITT_core_content_framework_.pdf

Kaiser, H. F. (1970). A second generation little jiffy. *Psychometrika, 35*(4), 401–415. https://doi.org/10.1007/BF02291817

Kang, S.H., McDermott, K., & Roediger, H.L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*, 528 - 558. https://doi.org/10.1080/09541440601056620

Karpicke, J.D., Blunt, J.R. & Sumeracki, M.A. (2016) Retrieval-based learning: Positive effects if retrieval practice in elementary school children. *Frontiers in Psychology 7* (104). https://doi.org/10.3389/fpsyg.2016.00350

Karpicke, J.D., Blunt, J.R., Smith, M.A., & Karpicke, S.S. (2014). Retrieval-based learning: The need for guided retrieval in elementary school students. *Journal of applied research in memory and cognition* 198-206 https://doi.org/10.1016/j.jarmac.2014.07.008

Karpicke, J.D., & Roediger, H.K. (2008). The critical importance of retrieval for learning. *Science 319*, 966-968. https://doi.org/10.1126/science.1152408

Karpicke, J.D., & Roediger H.L. (2007a). Expanding retrieval practice promotes short term retention but equally spaced retrieval enhances long term retention. *Journal of Experimental Psychology: Learning, Memory and Cognition, 33*(4), 704-719. https://doi.org/10.1037/0278-7393.33.4.704

Karpicke, J.D., & Roediger H.L. (2007b). Repeated retrieval during learning is the key to long-term retention. *Journal of memory and language 57*(2), 151-162 https://doi.org/10.1016/j.jml.2006.09.004

Kivikangas, J. M., Chanel, G., Cowley, B., Ekman, I., Salminen, M., Järvelä, S., & Ravaja, N. (2011). A review of the use of psychophysiological methods in game research. *Journal of Gaming and Virtual Worlds, 3*(3), 181–199. https://doi.org/10.1386/jgvw.3.3.181_1

Kliewer, W., Lepore, S. J., Farrell, A. D., Allison, K. W., Meyer, A. L., Sullivan, T. N., & Greene, A. Y. (2011). A school-based expressive writing intervention for at-risk urban adolescents' aggressive behavior and emotional lability. *Journal of clinical child and adolescent psychology: the official journal for the Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53*, *40*(5), 693–705. https://doi.org/10.1080/15374416.2011.597092

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. (4th ed.). Guilford publications.

Kline, P. (1994). *An Easy Guide to Factor Analysis.* Routledge. https://doi.org/10.4324/9781315788135

Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & cognition*, *34*(5), 959–972. https://doi.org/10.3758/bf03193244

Koriat, A., & Bjork, R.A. (2006). Mending metacognitive illusions: A comparison of mnemonic-based and theory-based procedures. *Journal of Experimental Psychology: Learning, Memory and Cognition 2006, Vol. 32*(5),1133-1145. https://doi.org/10.1037/0278-7393.32.5.1133

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(1), 283-294. https://doi.org/10.1037/a0037850

Kornell, N., & Bjork, R.A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General 138*(4) https://doi.org/10.1037/a0017350

Kornell, N., Hays, M.J., & Bjork, R.A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory and Cognition,35* (4) 989-998. https://doi.org/10.1037/a0015729

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction?" *Psychological Science, 19*(6), 585-592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14*(2), 219–224. https://doi.org/10.3758/BF03194055

Kreft, I., & de Leeuw, J. (1998). *Introducing statistical methods. Introducing multilevel modeling.* Sage Publications, Inc. https://doi.org/10.4135/9781849209366

Lang, J. W. B., & Lang, J. (2010). Priming Competence Diminishes the Link Between Cognitive Test Anxiety and Test Performance: Implications for the Interpretation of Test Scores. *Psychological Science*, *21*(6), 811–819. https://doi.org/10.1177/0956797610369492

LeBeau, R. T., Glenn, D., Liao, B., Wittchen, H. U., Beesdo-Baum, K., Ollendick, T., & Craske, M. G. (2010). Specific phobia: a review of DSM-IV specific phobia and preliminary recommendations for DSM-V. *Depression and anxiety*, *27*(2), 148–167. https://doi.org/10.1002/da.20655

Lee, N., Cadogan, J.W. & Chamberlain, L. The MIMIC model and formative variables: problems and solutions. *AMS Rev 3*, 3–17 (2013). https://doi.org/10.1007/s13162-013-0033-1

Lendrum, A. & Humphrey, N. (2012) The importance of studying the implementation of intervention in school settings. *Oxford Review of Education 38* (5) 635-652 https://doi.org/10.1080/03054985.2012.734800

Lepore, S. J., Greenberg, M. A., Bruno, M., & Smyth, J. M. (2002). *Expressive writing and health: Self-regulation of emotion-related experience, physiology, and behavior.* In S. J. Lepore & J. M. Smyth (Eds.), *The writing cure: How expressive writing promotes health and emotional well-being* (p. 99–117). American Psychological Association. https://doi.org/10.1037/10451-005

Lepore, S. J., & Smyth, J. M. (Eds.). (2002). *The writing cure: How expressive writing promotes health and emotional well-being.* American Psychological Association. https://doi.org/10.1037/10451-000

Levene, H. 1960. Robust tests for equality of variances. In: *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin, éds. Stanford University Press.

Liebert, R., & Morris, L. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological reports, 20,* 975-978, https://doi.org/10.2466/pr0.1967.20.3.975

Likert, R. (1932). "A Technique for the Measurement of Attitudes". *Archives of Psychology 22 140* 55.

Lipowski, S. L., Pyc, M. A., Dunlosky, J., & Rawson, K. A. (2014). Establishing and explaining the testing effect in free recall for young children. *Developmental Psychology, 50*(4), 994–1000. https://doi.org/10.1037/a0035202

Little, J.L., & Bjork, E.L. (2016) Multiple-choice pretesting potentiates learning of related information. *Mem Cogn 44*, 1085–1101. https://doi.org/10.3758/s13421-016-0621-z

Lohbeck, A., Nitkowski, D., & Petermann, F. (2016). A control-value theory approach: Relationships between academic self-concept, interest, and test anxiety in elementary school children *Child & Youth Care Forum, 45*(6) 887-904, https://psycnet.apa.org/doi/10.1007/s10566-016-9362-1

Lowe, P. A., Lee, S. W., Witteborg, K. M., Prichard, K. W., Luhr, M. E., Cullinan, C. M., Mildren, B. A., Raad, J. M., Cornelius, R. A., & Janik, M. (2008). The Test Anxiety Inventory for Children and Adolescents (TAICA): Examination of the Psychometric Properties of a New Multidimensional Measure of Test Anxiety Among Elementary and Secondary School Students. *Journal of Psychoeducational Assessment*, *26*(3), 215–230. https://doi.org/10.1177/0734282907303760

Marchand, E., Stice, E., Rohde, P., & Becker, C. B. (2011). Moving from efficacy to effectiveness trials in prevention research. *Behaviour research and therapy*, *49*(1), 32–41. https://doi.org/10.1016/j.brat.2010.10.008

Marsh, E. J., Agarwal, P. K., & Roediger, H. L. III. (2009). Memorial consequences of answering SAT II questions. *Journal of Experimental Psychology: Applied, 15*(1), 1–11. https://doi.org/10.1037/a0014721

Marsh, H.W., Hau, K.T., Balla, J.R. & Grayson, D (1998) Is More Ever Too Much? The Number of Indicators per Factor in Confirmatory Factor Analysis, *Multivariate Behavioral Research,33:2,* 181-220, https://doi.org/10.1207/s15327906mbr3302_1

Martin, A. (2002). Motivation and Academic Resilience: Developing a Model for Student Enhancement. *Australian Journal of Education*, *46*(1), 34–49. https://doi.org/10.1177/000494410204600104

Martin, A.J. (2007). Examining a multidimensional model of student motivation and engagement using a construct validation approach. *British Journal of Educational Psychology*, *77*, 413-440. https://doi.org/10.1348/000709906X118036

Martin, A. J. (2013). Academic buoyancy and academic resilience: Exploring 'everyday' and 'classic' resilience in the face of academic adversity. *School Psychology International, 34*(5), 488–500. https://doi.org/10.1177/0143034312472759

Martin, A. J., Colmar, S. H., Davey, L. A., & Marsh, H. W. (2010). Longitudinal modelling of academic buoyancy and motivation: do the '5Cs' hold up over time? *The British journal of educational psychology*, *80*(3), 473–496. https://doi.org/10.1348/000709910X486 376

Martin, A. J., & Marsh, H. W. (2008). Academic buoyancy: Towards an understanding of students' everyday academic resilience. *Journal of school psychology*, *46*(1), 53–83. https://doi.org/10.1016/j.jsp.2007.01.002

Matthews, G., Hillyard, E.J., & Campbell, S. (1999). Metacognition and maladaptive coping as components of test anxiety. *Clinical Psychology & Psychotherapy, 6*, 111-125. https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1099-0879(199905)6:2%3C111::AID-CPP192%3E3.0.CO;2-4

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27*(3), 360–372. https://doi.org/10.1002/acp.2914

McIntyre, L. L., Gresham, F. M., DiGennaro, F. D., & Reed, D. D. (2007). Treatment integrity of school-based interventions with children in the journal of applied behavior analysis 1991-2005. *Journal of applied behavior analysis*, *40*(4), 659–672. https://doi.org/10.1901/jaba.2007.659-672

McKeachie, W. J. (1951). Anxiety in the college classroom. *The Journal of Educational Research, 45,* 153–160.

McKenney, S., & Reeves, T. C. (2013). Systematic Review of Design-Based Research Progress: Is a Little Knowledge a Dangerous Thing? *Educational Researcher*, *42*(2), 97–100. https://doi.org/10.3102/0013189X12463781

McNamara D. S. (2010). Strategies to read and learn: overcoming learning by consumption. *Medical education*, *44*(4), 340–346. https://pubmed.ncbi.nlm.nih.gov/20236240/

Metcalfe, J. (2017). Learning from errors. *Annual review of psychology*, *68*, 465-489. https://doi.org/10.1146/annurev-psych-010416-044022

Metcalfe, J., Kornell, N. & Finn, B. (2009) Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition 37***,** 1077–1087 (2009). https://doi.org/10.3758/MC.37.8.1077

Mertens, D.M. (2015) *Research and Evaluation in Education and Psychology.* Sage Thousand Oaks

Murayama, K., Goetz, T., Malmerg, L.E., Pekrun, R., Tanaka, A., & Martin, A.J. (2017) *Within-person analysis in educational psychology: Importance and illustrations*. In D.W. Putwain & K. Smart (Eds.) *The role of competence beliefs in teaching and learning* (pp.71-88). The British Psychological Society.

Murphy, S. L., & Gutman, S. A. (2012). Intervention fidelity: a necessary aspect of intervention effectiveness studies. *The American journal of occupational therapy: official publication of the American Occupational Therapy Association*, *66*(4), 387–388. https://doi.org/10.5014/ajot.2010.005405

Naveh-Benjamin, M. (1991). A comparison of training programs intended for different types of test-anxious students: Further support for an information-processing model. *Journal of Educational Psychology, 83*(1), 134–139.https://doi.org/10.1037/0022-0663.83.1.134

Nelson, T.O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & Arthur P. Shimamura (Eds.), *Metacognition knowing about knowing* (pp.1-23). A Bradford Book. The MIT Press. https://doi.org/10.7551/mitpress/4561.001.0001

Nordman, E. (2020) MSc Conversion in Psychological Studies, https://psyteachr.github.io/msc-conv-odl/

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15*, 625-632. https://doi.org/10.1007/s10459-010-9222-y

Nunnally, J. C. (1978). An overview of psychological measurement. *Clinical diagnosis of mental disorders*, 97-146. https://doi.org/10.1007/978-1-4684-2490-4_4

OSF (Open Science Framework, https://osf.io/j4dn2/)

Owens,M., Stevenson, J., Norgate, R., & Hadwin, J.A. (2008)Processing efficiency theory in children: Working memory as a mediator between trait anxiety and academic performance, *Anxiety, Stress & Coping, 21:4*, 417-430. https://doi.org/10.1080/10615800701847823

Owens, M., Stevenson, J., Hadwin, J.A., & Norgate, R. (2012). When does anxiety help or hinder cognitive test performance? The role of working memory capacity. *British Journal of Psychology 105* 92-10. https://doi.org/10.1111/bjop.12009

Pan, S.C., Bjork, R.A. (in press) Acquiring a mental model of learning: Towards an owner's manual. In A. Wagner and M. Kahana (Eds.), *Oxford Handbook of Learning and Memory: Foundations and Applications*. Oxford University Press https://bjorklab.psych.ucla.edu/publication/pan-s-c-and-bjork-r-a-in-press-acquiring-an-accurate-mental-model-of-learning-towards-an-owners-manual-in-a-wagner-and-m-kahana-eds-oxford-handbook-of-learning-memory/

Pan, S.C., Rickard, T.C., & Bjork, R.A. (2021) Does spelling still matter – and if so how should it be taught? Perspectives from contemporary and historical research. *Educational Psychology Review.* https://doi.org/10.1007/s10648-021-09611-y

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological bulletin*, *144*(7), 710–756. https://doi.org/10.1037/bul0000151

Panhwar, A.H., Ansari, S., & Shah, A. (2017). Post-positivism: an effective paradigm for social and educational research. *International Research Journal of Arts and Humanities. Vol.45* (45) Sindh University Press. https://sujo-old.usindh.edu.pk/index.php/IRJAH/article/view/3371

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational psychology review*, *18*(4), 315-341. https://doi.org/10.1007/s10648-006-9029-9

Phillips, D. C., & Burbules, N. C. (2000). *Philosophy, theory and educational research. Postpositivism and educational research.* Rowman & Littlefield. https://psycnet.apa.org/record/2000-08253-000

Pillinger, R. (n.d) *Random intercept models.* University of Bristol Centre for multilevel linear modelling. http://www.bristol.ac.uk/cmm/learning/videos/random-intercepts.html

Pintado, I. S., Sánchez-Mateos, J. D., & Escolar-Llamazares, M. C. (2016). A stress inoculation program to cope with test anxiety: Differential efficacy as a function of worry or emotionality. *Avances en Psicología Latinoamericana, 34*(1), 3–18. https://psycnet.apa.org/doi/10.12804/apl34.1.2016.01

Putnick, D.L., Bornstein, M.H. (2016) Measurement invariance conventions and reporting: The state of the art and future directions for psychological research, *Developmental Review,Volume 41,* 71-90 https://doi.org/10.1016/j.dr.2016.06.004

Putwain, D.W., Gallard, DC, Beaumont, J, Loderer, K and von der Embse, N (2021) Does Test Anxiety Predispose Poor School-Related Wellbeing and Enhanced Risk of Emotion Disorders? *Cognitive Therapy and Research.* LJMU Research online. https://researchonline.ljmu.ac.uk/id/eprint/14279/

Putwain, D. W. (2020) *Examination Pressures on Children and Young People: Are They Taken Seriously Enough? A Provocation Paper*. The British Academy. https://medium.com/reframing-childhood-past-and-present/examination-pressures-on-children-and-young-people-are-they-taken-seriously-enough-e274b9595d4

Putwain, D. W., von der Embse, N. P., Rainbird, E. C., & West, G. (2020). The development and validation of a new Multidimensional Test Anxiety Scale (MTAS). *European Journal of Psychological Assessment.* Advance online publication. https://doi.org/10.1027/1015-5759/a000604

Putwain, D.W. (2018) An examination of the self-referent executive processing model of test anxiety: control, emotional regulation, self-handicapping, and examination performance. *European Journal of Psychology and Education,* 34, 341-358 https://doi.org/10.1007/s10212-018-0383-z

Putwain, D.W., & Aveyard, B (2018). Is perceived control a critical factor in understanding the negative relationship between cognitive test anxiety and examination performance? *School Psychology Quarterly 33* (1) pp 65-74. Advance online publication. https://doi.org/10.1037/spq0000183

Putwain, D.W., & Pescod, M. (2018). Is reducing uncertain control the key to successful test anxiety intervention for secondary school students? Findings from a randomized control trial. *School Psychology Quarterly Vol 33* (2) 283-292. https://doi.org/10.1037/spq0000228

Putwain, D.W., Symes, W. (2018) Does increased effort compensate for performance debilitating test anxiety? *School Psychology Quarterly*. https://psycnet.apa.org/doi/10.1037/spq0000236

Putwain, D. W., Daly, A. L., Chamberlain, S., & Sadreddini, S. (2016). "'Sink or swim': Buoyancy and coping in the cognitive test anxiety—Academic performance relationship". *Educational Psychology, 36*(10), 1807–1825. https://doi.org/10.1080/01443410.2015.1066493

Putwain, D.W., Remedios, R., & Symes, W. (2016). The appraisal of fear appeals as threatening or challenging: frequency of use, academic self-efficacy and subjective value. *Educational Psychology Vol 36.* https://doi.org/10.1080/01443410.2014.963028

Putwain, D. W., Symes, W., & Remedios, R. (2016). The impact of fear appeals on subjective-task value and academic self-efficacy: The role of appraisal. *Learning and Individual Differences, 51,* 307–313. https://doi.org/10.1016/j.lindif.2016.08.042

Putwain, D. W., Daly, A. L., Chamberlain, S., & Sadreddini, S. (2015). Academically buoyant students are less anxious about and perform better in high-stakes examinations. *The British journal of educational psychology*, *85*(3), 247–263. https://doi.org/10.1111/bjep.12068

Putwain, D. W., Shah, J., & Lewis, R. (2014). Performance-evaluation threat does not adversely affect verbal working memory in high test-anxious persons. *Journal of Cognitive Education and Psychology, 13*(1), 120–136. https://doi.org/10.1891/1945-8959.13.1.120

Putwain, D.W., Chamberlain, S., Daly, A.L., & Sadreddini, S. (2014). Reducing test anxiety among school-aged adolescents: a field experiment. *Educational Psychology in Practice 30.4*, 420-44, https://doi.org/10.1080/02667363.2014.964392

Putwain, D.W., Daly, A.L. (2014) Test anxiety prevalence and gender differences in a sample of English secondary school students, *Educational Studies*, *40:5*, 554- 570, https://doi.org/10.1080/03055698.2014.953914

Putwain, D. W., Nicholson, L. J., Connors, L., & Woods, K. (2013). Resilient children are less test anxious and perform better in tests at the end of primary schooling. *Learning and Individual Differences, 28,* 41–46. https://doi.org/10.1016/j.lindif.2013.09.010

Putwain, D. W., & Best, N. (2011). Fear appeals in the primary classroom: Effects on test anxiety and test grade. *Learning and Individual Differences, 21*(5), 580–584. https://psycnet.apa.org/doi/10.1016/j.lindif.2011.07.007

Putwain, D.W. (2008a). Deconstructing test anxiety *Journal of Emotional and Behavioural Difficulties 13*, 141-155. https://doi.org/10.1080/13632750802027713

Putwain, D. (2008). Examination stress and test anxiety. *The Psychologist, 21*(12), 1026–1029. https://thepsychologist.bps.org.uk/volume-21/edition-12/examination-stress-and-test-anxiety

Putwain, D. (2008). Do examinations stakes moderate the test anxiety-examination performance relationship? *Educational Psychology, 28*(2), 109–118. https://doi.org/10.1080/01443410701452264

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*(6002), 335. https://doi.org/10.1126/science.1191465

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Ramirez, G., & Beilock, S. L. (2011). Writing about testing worries boosts exam performance in the classroom. *Science (New York, N.Y.)*, *331*(6014), 211–213. https://doi.org/10.1126/science.1199427

Rasbash, J. (n.d.) *What are multilevel models and why should I use them?* University of Bristol Centre for multilevel modelling http://www.bristol.ac.uk/cmm/learning/multilevel-models/what-why.html)

Ritchie, S. J., Della Sala, S., & McIntosh, R. D. (2013). Retrieval practice, with or without mind mapping, boosts fact learning in primary school children. *PloS one*, *8*(11), e78976. https://doi.org/10.1371/journal.pone.0078976

Rickard, T. C., & Pan, S. C. (2018). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review, 25*(3), 847–869. https://doi.org/10.3758/s13423-017-1298-4

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*(4), 382–395. https://doi.org/10.1037/a0026252

Roediger, H. L., & Karpicke, J. D. (2011). Intricacies of spaced retrieval: A resolution. In A. S. Benjamin (Ed.), *Successful Remembering and Successful Forgetting: Essays in Honor of Robert A. Bjork* (pp. 23-48). New York: Psychology Press.

Roediger, H.L., Agarwal, P.K., Kang, S.H.K., & Marsh, E.J. (2010). *Benefits of testing memory: Best practices and boundary conditions.* In G.M. Davies & D.B. Wright (Eds.), *Current issues in memory. Current issues in applied memory research* (p.13-49). Psychology Press.

Roediger, H.L., & Karpicke J.D. (2008). The critical importance of retrieval for learning. *SCIENCE Vol 319* 966-968. https://doi.org/10.1126/science.1152408

Roediger, H.L., & Karpicke J.D. (2006b). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science 17*(3) 249-255. https://doi.org/10.1111%2Fj.1467-9280.2006.01693.x

Roediger, H.L., & Karpicke J.D. (2006a). The power of testing memory. Basic research and implications for educational practice. *Perspectives on Psychological Science 1*(3), 181-210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of experimental psychology. Learning, memory, and cognition*, *31*(5), 1155–1159. https://doi.org/10.1037/0278-7393.31.5.1155

Rosseel, Y. (2012). "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, 48(2), 1–36. http://www.jstatsoft.org/v48/i02/

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Sarason, I., Sarason, B., Keefe, D., Hayes, B., & Shearin, E. (1986). Cognitive interference: Situational determinants and traitlike characteristics. *Journal of Personality and Social Psychology*, *51*, 215–226. https://insights.ovid.com/personality-social-psychology/jpspy/1986/07/000/cognitive-interference-situational-determinants/21/00005205

Sarason, I.G. (1984). Stress, anxiety and cognitive interference: Reactions to tests. *Journal of Personality and Social Psychology, 46*(4) 929-938. https://doi.org/10.1037/0022-3514.46.4.929

Sarason, I. G. (1972). Test anxiety and the model who fails. *Journal of Personality and Social Psychology, 22*(3), 410–413. https://doi.org/10.1037/h0032893

Sarason, I.G. (1958) The effects on verbal learning of anxiety, reassurance and meaningfulness of material. *Journal of Experimental Psychology, 56* (6)473-477. https://doi.org/10.1037/h0049335

Sarason, S. B., Davidson, K. S., Lighthall, F. F., Waite, R. R., & Ruebush, B. K. (1960). *Anxiety in elementary school children: A report of research.* John Wiley & Sons Inc. https://doi.org/10.1037/14349-000

Sarason, S., Davidson, K., Lighthall, F., & Waite, R. (1958). A Test Anxiety Scale for Children. *Child Development, 29(1),* 105-113. https://doi.org/10.2307/1126274

Sarnoff, I., Lighthall, F., Waite, R., Davidson, K., & Sarason, S. (1958). A cross-cultural study of anxiety among American and English school children. *Journal of Educational Psychology, 49*(3), 129–136. https://doi.org/10.1037/h0043962

Schneider BA, Avivi-Reich M and Mozuraitis M (2015) A cautionary note on the use of the Analysis of Covariance (ANCOVA) in classification designs with and without within-subject factors. *Front. Psychol. 6:474.* https://doi.org/10.3389/fpsyg.2015.00474

Scammacca, N. K., Roberts, G., Vaughn, S., & Stuebing, K. K. (2015). A Meta-Analysis of Interventions for Struggling Readers in Grades 4-12: 1980-2011. *Journal of learning disabilities*, *48*(4), 369–390. https://doi.org/10.1177/0022219413504995

Schunk, D.H. (1989) Self-efficacy and cognitive skill in learning. In C. Ames and R.Ames (Eds.), *Research on motivation in education. Vol.3.Goals and Cognitions* (pp.13-44) Academic

Schunk, D. H., & Pajares, F. (2005). *Competence Perceptions and Academic Functioning.* In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (p. 85–104). Guilford Publications.

Schutz, P. A., Davis, H. A., Decuir-Gunby, J. T., & Tillman, D. (2014). *Regulating emotions related to testing.* In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *Educational psychology handbook series. International handbook of emotions in education* (p. 348–367). Routledge/Taylor & Francis Group.

Schutz, P. A., Benson, J., & Decuir-Gunby, J. T. (2008). Approach/avoidance motives, test emotions, and emotional regulation related to testing. *Anxiety, stress, and coping*, *21*(3), 263–281. https://doi.org/10.1080/10615800701787672

Schutz, P. A., Davis, H. A., & Schwanenflugel, P. J. (2002). Organization of concepts relevant to emotions and their regulation during test taking. *Journal of Experimental Education, 70*(4), 316–342. https://doi.org/10.1080/00220970209599511

Schutz, P.A. & Davis, H.A. (2000) Emotions and Self-Regulation During Test Taking, *Educational Psychologist, 35:4,* 243-256, https://10.1207/S15326985EP3504_03

Schwarzer, R. & Jerusalem, M. (1992) Advances in anxiety theory: A cognitive process approach. In K.A. Hagtvet & T/B. Johnsen (Eds.), *Advances in test anxiety research (Vol. 7*) 2-31. Swets & Zeitlinger

Schwarzer, R. (1990). Current Trends in Anxiety Research. In P. J. D. Drenth, J. A. Sergeant, & R. J. Takens (Eds.), *European Perspectives in Psychology* (pp. 225-244). Wiley.

Segool, N.K., Carlson, J.S., Goforth, A.N., Von der Embse, N., & Barterian, J.A., (2013). Heightened Test Anxiety Among Young Pupils: Elementary School Students' Anxious Responses to High Stakes Testing *Psychology in the Schools Vol. 50* (5). https://doi.org/10.1002/pits.21689

Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research, 4*(1), 27–41. https://doi.org/10.1080/08917779108248762

Shapiro, E. S. (1996). Academic skills problems: Direct assessment and intervention. Guilford.

Shapiro, S.S. & Wilk, M.B. (1965) An analysis of variance test, for normality and complete samples *Biometrika, 52,* pp. 592- 611. https://doi.org/10.2307/2333709

Shen, L., Yang, L., Zhang, J., & Zhang, M. (2018). Benefits of expressive writing in reducing test anxiety: A randomized controlled trial in Chinese samples. *PloS one*, *13*(2), e0191779. https://doi.org/10.1371/journal.pone.0191779

Shuttleworth. M. (Nov 3, 2009). Pretest-Posttest Designs. Explorable.com: https://explorable.com/pretest-posttest-designs

Singal, A. G., Higgins, P. D., & Waljee, A. K. (2014). A primer on effectiveness and efficacy trials. *Clinical and translational gastroenterology*, *5*(1), e45. https://doi.org/10.1038/ctg.2013.13

Smith, A. M., Floerke, V. A., & Thomas, A. K. (2016). Retrieval practice protects memory against acute stress. *Science (New York, N.Y.)*, *354*(6315), 1046–1048. https://doi.org/10.1126/science.aah5067

Smith, S. M., Glenberg, A., & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition, 6*(4), 342–353. https://doi.org/10.3758/BF03197465

Smith, R.E. (1989) Effects of Coping Skills Training on Generalized Self-Efficacy and Locus of Control. *Journal of Personality and Social Psychology, 56,* 228-233. https://doi.org/10.1037/0022-3514.56.2.228

Smith, R. J., Arnkoff, D. B., & Wright, T. L. (1990). Test anxiety and academic competence: A comparison of alternative models. *Journal of Counseling Psychology, 37*(3), 313–321. https://doi.org/10.1037/0022-0167.37.3.313

Sommer, M., & Arendasy, M. E. (2014). Comparing different explanations of the effect of test anxiety on respondents' test scores. *Intelligence, 42,* 115–127. https://doi.org/10.1016/j.intell.2013.11.003

Sperling, R. A., Reeves, P. M., Gervais, A. L., & Sloan, J. O. (2017). Development and Validation of the Self-Regulated Learning Test Anxiety Scale. *Psychology*, *8*(14), 2295. https://doi.org/10.4236/psych.2017.814145

Spielberger, C.D., & Vagg, P.R. (1995). Test anxiety. A Transactional Process Model. In C.D. Spielberger & P.R. Vagg (Eds.), *Test anxiety, theory, assessment and treatment*. (pp.3-14) Taylor & Francis.

Spielberger, C.D., Anton, W.D., & Bedell, J. (1976). *The nature and treatment of test anxiety.* In M. Zuckerman & C.D. Spielberger (Eds.), *Emotions and anxiety*, (pp. 317-341) Lawrence Erlbaum Associates Inc.

Spielberger, C.D., Edwards, C.D., Lushene, R.E., Montuori, J., & Platzek, D. (1973). *STAIC preliminary manual for the State-Trait Anxiety Inventory for Children*. Consulting Psychologists Press

Steinmayr, R., Crede, J., McElvany, N., & Wirthwein, L. (2016). Subjective Well-Being, Test Anxiety, Academic Achievement: Testing for Reciprocal Effects. *Frontiers in psychology*, *6*, 1994. https://doi.org/10.3389/fpsyg.2015.01994

Stowell, J. R., Tumminaro, T., & Attarwala, M. (2008). Moderating effects of coping on the relationship between test anxiety and negative mood. *Stress and Health: Journal of the International Society for the Investigation of Stress*, *24*(4), 313-321. https://doi.org/10.1002/smi.1186

Sullivan, G. M., & Artino, A. R., Jr (2013). Analyzing and interpreting data from likert-type scales. *Journal of graduate medical education*, *5*(4), 541–542. https://doi.org/10.4300/JGME-5-4-18

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing Three Sources of Misresponse to Reversed Likert Items. *Journal of Marketing Research*, *45*(1), 116–131. https://doi.org/10.1509%2Fjmkr.45.1.116

Swann, J. (2003) How Science can Contribute to the Improvement of Educational Practice, *Oxford Review of Education, 29:2,* 253-268, https://doi.org/10.1080/0305498032000080710

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Allyn & Bacon/Pearson Education

TED (2014, November) Carol Dweck: The power of yet [video] YouTube https://www.ted.com/talks/carol_dweck_the_power_of_believing_that_you_can_improve?language=ry#t-2903

Thurstone, L. L. (1947). Multiple-factor analysis; a development and expansion of The Vectors of Mind. University of Chicago Press.

Tobias, S. (1985). Test anxiety: Interference, defective skills, and cognitive capacity. *Educational Psychologist, 20*(3) 135-142. https://doi.org/10.1207/s15326985ep2003_3

Tsapali, M. (2019). *Effects of different learning environments on late primary school students' Decision-making Competence in Socio-Scientific Issues* (Doctoral thesis). https://doi.org/10.17863/CAM.54222

Tse, C., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working memory capacity. *Journal of Experimental Psychology*: *Vol.18,* (3), 253-264. https://doi.org/10.1037/a0029190

Tullis, J.G., Finley, J.R., & Benjamin, A.S. (2012). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Mem. Cogn 41*:429-442. https://doi.org/10.3758/s13421-012-0274-5

Twisk, J. (2006). *Applied Multilevel Analysis: A Practical Guide for Medical Researcher*s (Practical Guides to Biostatistics and Epidemiology). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511610806

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of reverse wording of questionnaire items: let's learn from cows in the rain. *PloS one*, *8*(7), e68967. https://doi.org/10.1371/journal.pone.0068967

von der Embse, N/P., Putwain., D.W., Francis, G. (2021) Interpretation and use of the multidimensional test anxiety scale (MTAS) *American Psychological Association.* https://researchonline.ljmu.ac.uk/id/eprint/14280

von der Embse, N., Jester, D., Devlina, R., Post, J. (2018) Test anxiety effects, predictors, and correlates: A 30-year meta-analytic review. *Journal of Affective Disorders* 227, 483–493. https://doi.org/10.1016/j.jad.2017.11.048

von der Embse, N., Barterian, J., & Segool, N. (2013). Test anxiety interventions for children and adolescents: A systematic review of treatment studies from 2000-2010. *Psychology in the Schools, 50*, 57-71. https://doi.org/10.1002/pits.21660

Wacker, D. P., Steege, M. W., Northup, J., Sasso, G., Berg, W., Reimers, T., Cooper, L., Cigrand, K., & Donn, L. (1990). A component analysis of functional communication training across three topographies of severe behavior problems. *Journal of applied behavior analysis*, *23*(4), 417–429. https://doi.org/10.1901/jaba.1990.23-417

Taiyun, W. & Viliam, S. (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot

Weems, C.F., Scott, B.G., Graham, R.A., Banks, D.M., Russell, D.J., Taylor, L.K., Cannon, M.F., Varela, R.E., Scheeringa, M.A., Perry, A.M., & Marino, R.C. (2015). Fitting anxious emotion-focused intervention into the ecology of schools: Results from a test anxiety program evaluation. *Prevention Science, 16*(2), 200-210. https://doi.org/10.1007/s11121-014-0491-1

Weems, C. F., Scott, B. G., Taylor, L. K., Cannon, M. F., Romano, D. M., Perry, A. M., & Triplett, V. (2010). Test anxiety prevention and intervention programs in schools: Program development and rationale. *School Mental Health: A Multidisciplinary Research and Practice Journal, 2*(2), 62–71. https://doi.org/10.1007/s12310-010-9032-7

Weems, C.F., Taylor, L.K., Costa, N.M., Marks, A.B., Romano, D.M., Verrett, S.L., Darlene Brown, D.M. (2009) Effect of a school-based test anxiety intervention in ethnic minority youth exposed to Hurricane Katrina, *Journal of Applied Developmental Psychology, Volume 30, Issue 3,* 218-226. https://doi.org/10.1016/j.appdev.2008.11.005

Weems, G. H., Onwuegbuzie, A. J., Schreiber, J. B., & Eggers, S. J. (2003). Characteristics of Respondents Who Respond Differently to Positively and Negatively Worded Items on Rating Scales. *Assessment & Evaluation in Higher Education, 28*(6), 587–607. https://doi.org/10.1080/0260293032000130234

Weiner, B. (2010). The development of an attribution-based theory of motivation: A history of ideas. *Educational Psychologist, 45*(1), 28-36. https://doi.org/10.1080/00461520903433596

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning & Verbal Behavior, 16*(4), 465–478. https://doi.org/10.1016/S0022-5371(77)80040-6

World Health Organization (1992). *International classification of diseases and related health problems* (10th rev., ICD-10). Geneva, Switzerland: Author.

Wilson, E. O. (1998). Consilience among the great branches of learning. *Daedalus*, *127*(1), 131-149.

Willingham, D.T. (2007) Ask the Cognitive Scientist: The usefulness of brief instruction in reading comprehension strategies. *American Educator: 2007; 30 (4):* 39-45.50. https://www.aft.org/sites/default/files/periodicals/CogSci.pdf

Willingham, D.T. (2012, April 30th) Collateral damage of excessive reading comprehension strategy instruction [blog post], http://www.danielwillingham.com/daniel-willingham-science-and-education-blog/collateral-damage-of-reading-comprehension-strategy-instruction

Williams, Robert L. & Bush, Vivian J. & Park, Soo-Hee. & Malone, L. Yvonne. & Jessup, Kary. (2001). *Work Ethic Scale for Middle School Students*. (ED452257). ERIC. https://eric.ed.gov/?id=ED452257

Wilson, E. (Ed.). (2017). *School-based research: A guide for education students*. Sage.

Wine, J. (1971). Test anxiety and direction of attention. *Psychological Bulletin, 76*(2) 92-104. https://doi.org/10.1037/h0031332

Wine, J. D. (1982). Evaluation anxiety: A cognitive-attentional construct. *Series in Clinical & Community Psychology: Achievement, Stress, & Anxiety,* 207–219.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and psychological measurement*, *76*(6), 913–934. https://doi.org/10.1177/0013164413495237

Wong, S.S. (2008). The Relations of Cognitive Triad, Dysfunctional Attitudes, Automatic Thoughts, and Irrational Beliefs with Test Anxiety. *Current Psychology, 27*, 177-191. https://doi.org/10.1007/s12144-008-9033-y

Worthington, R. L., & Whittaker, T. A. (2006). Scale Development Research: A Content Analysis and Recommendations for Best Practices. *The Counseling Psychologist, 34*(6), 806–838. https://doi.org/10.1177/0011000006288127

Wren, D.G., & Benson, J (2004). Measuring test anxiety in students: Scale development and internal construct validation *Anxiety, Stress and Coping, 17.3,* 227-240, https://doi.org/10.1080/10615800412331292606

Yan, V.X., Bjork, E.L., & Bjork, R.A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology*: General 2016. Vol 145 (7), 918-933. https://doi.org/10.1037/xge0000177

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin.* Advance online publication. https://doi.org/10.1037/bul0000309

Yang, C., Bukuan, S., Potts, R., Rongjun, Y., Liang, L., Shanks, R. (2020) Do working memory capacity and test anxiety modulate the beneficial effects of testing on new learning?*Journal of Experimental Psychology: Applied, Vol 26(4),* 724-738 http://dx.doi.org/10.1037/xap0000278

Yeager, D. S., Trzesniewski, K. H., & Dweck, C. S. (2013). An implicit theories of personality intervention reduces adolescent aggression in response to victimization and exclusion. *Child development*, *84*(3), 970–988. https://doi.org/10.1111/cdev.12003

Yeager, D.S., & Walton, G.M. (2011) Social-Psychological Interventions in Education: They're Not Magic *Review of Educational Research*, *81*(2), 267 - 301. https://doi.org/10.3102/0034654311405999

Yerkes, R.M. & Dodson, J.D. (1908) The relation of strength of stimulus to rapidity of habit Formation. *Journal of Comparative Neurology and Psychology 18 (5)* https://doi.org/10.1002/cne.920180503

Yeo, L. S., Goh, V. G., & Liem, G. A. D. (2016). School-based intervention for test anxiety. *Child & Youth Care Forum, 45*(1), 1–17. https://doi.org/10.1007/s10566-015-9314-1

Young, J. E., Klosko, J. S., & Weishaar, M. E. (2003). *Schema therapy: A practitioner's guide.* Guilford Press.

Zeidner, M. (2014). Anxiety in education. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (pp. 265-285) Routledge. Taylor & Francis Group

Zeidner, M (2007). Test anxiety: Conceptions, findings, conclusions. In P. Schutz & R.Pekrun (Eds.), *Emotion in education* (pp. 165-184). Sage.

Zeidner, M., & Matthews, G. (2005). Evaluation anxiety. Current theory and research. In A.J. Elliot & C.S. Dweck (Eds.), *Handbook of competence and motivation.* (pp. 141-160) The Guildford Press, New York and London.

Zeidner, M (1998) *Test anxiety. The state of the art.* Springer Science+Business

Zepeda, C.D., Martin, R.S., Butler, A.C.(2020) Motivational Strategies to Engage Learners in Desirable Difficulties, *Journal of Applied Research in Memory and Cognition ,Vol 9,* (4) 468-474, https://doi.org/10.1016/j.jarmac.2020.08.007

Zhang, X., & Savalei, V. (2016). Improving the Factor Structure of Psychological Scales: The Expanded Format as an Alternative to the Likert Scale Format. *Educational and Psychological Measurement*, *76*(3), 357–386. https://doi.org/10.1177/0013164415596421

Appendices

Appendix A – Questionnaires

Children's Test Attitude Questionnaire (English Version)

Please answer the questions using the following numbers:
1= almost never
2= some of the time
3=most of the time
4=almost always

Pupil initials----------
Boy /girl

1. While I am taking tests I wonder if I will pass--------
2. While I am taking tests my heart beats fast----------
3. While I am taking tests I look around the room--------
4. While I am taking tests I feel nervous----------
5. While I am taking tests I think I'm going to get a bad result---------
6. While I am taking tests it's hard for me to remember the answers--------
7. While I am taking tests I play with my pen /pencil----------
8. While I am taking tests my face feels hot----------
9. While I am taking tests I worry about failing----------
10. While I am taking tests my belly feels funny----------
11. While I am taking tests I worry about doing something wrong----------
12. While I am taking tests I check the time----------
13. While I am taking tests I think about what my result will be----------
14. While I am taking tests I find it hard to sit still----------
15. While I am taking tests I wonder if my answers are right---------
16. While I am taking tests I think I should have studied more---------
17. While I am taking tests my head hurts----------
18. While I am taking tests I look at other people----------
19. While I am taking tests I think most of my answers are wrong-------
20. While I am taking tests I feel warm--------
21. While I am taking tests I worry about how hard the test is----------
22. While I am taking tests I try to finish quickly----------
23. While I am taking tests my hand shakes----------
24. While I am taking tests I think about what will happen if I fail----------
25. While I am taking tests I have to go to the bathroom----------
26. While I am taking tests I tap my feet---------
27. While I am taking tests I think about how badly I am doing-------
28. While I am taking tests I feel scared-------
29. While I am taking tests I worry about what my parents will say--------
30. While I am taking tests I stare---------

Self-efficacy in Test Taking Questionnaire

HCAY (How Confident Are You?) Pupil initials----

How Confident Are You? Boy /girl

1= "I never do this".
2= "I do this sometimes."
3= "I do this many times"
4= "I always do this"

1. If I find a test hard, I keep trying.
2. If I find a test hard, I still feel confident in my ability to pass.
3. If I did badly on a test, I still feel positive about other tests.
4. If I find a test hard, I feel I will not pass.
5. When I take tests, I feel anxious about being in a different room to my usual classroom.
6. When I take tests, I feel anxious about the instructor.
7. If I find a test hard, I can't control negative thoughts about failing.
8. If I find a test hard, I start to doubt my ability to pass.
9. In tests, I panic if I see people finishing before me.
10. I can concentrate in tests, even if I find them difficult.
11. I enjoy taking tests.
12. I can remain cheerful about life, even if a test is hard.
13. I can manage in tests, even if I feel tired or unwell.
14. I understand that even if I find a test difficult, it is not the end of the world.
15. I don't worry about taking tests.

.

Appendix B

Teacher's Manual for the Intervention

Teacher's Manual for the 6 weeks Cognitive Intervention Based on Bjork's (1975) 'The testing effect'

By
Helen Barsham
hb479@cam.ac.uk

UNIVERSITY OF CAMBRIDGE
Faculty of Education



INSTRUCT Research Group
Psychology and Education
Web:
http://sites.google.com/site/instructlab/
PI: Dr Michelle Ellefson
Email:
Phone: 01223 767684

Contents

Introduction – P3

Instructions – P4

Test Attitude Questionnaire – P6 (pre)&7(post)

Self-efficacy Questionnaire- P8 (pre) &9 (post) called 'How Confident Are You?'

Teacher Questionnaire – P10

Questions for students (post-test) – P11

Cognition worksheets - P12-20

Sample permissions letter – P21

Supporting flyer -P22

Further Reading P 23-35 (not necessary but in case ….)

Introduction:

Thank you for agreeing to help in this exciting research, which aims to improve pupils' wellbeing.

In 2017-2018, I conducted a pilot study with a very small sample and the results were statistically significant. It made me determined to help more pupils in testing situations.

Your help will enable me to truly assess the impact of this intervention and to create an optimal intervention and to spread the word.

The intervention is aimed at 10-to11- year- olds who have to take a high stakes' exam, whether this assessment is a SATS exam or an entrance exam, 11+ etc. A weekly homework sheet (x6 weeks) followed by 15-20 minutes of class discussion (x6) is all it takes to help pupils to feel more self-efficacious in testing situations.

The research is based on Bjork's (1975) *desirable difficulty* of *the testing effect* and Bandura's (1997) work on self-efficacy and test anxiety theory (e.g., Zeidner, 1998). I designed the worksheets by reading the relevant authors and papers and putting these key messages into Year 6 pupil friendly language and creating a narrative, in which the authors themselves are the characters.

The rest, as they say, is history….

Instructions:

1. Administer the Children's Test Attitude Survey (marked pre-test) –takes about 15 mins. This should be done approximately 6 weeks before the high stakes' test to avoid any sense of priming. Collate questionnaires and hang on to them –put somewhere safe or email over -password protect.

2. At the same time as the CTAS (Children's Test Attitude Survey), administer the Self-efficacy Scale About Testing (How Confident Are You? HCAY) questionnaire (marked pre-test) –takes about 10 mins and keep somewhere safe or email password protect.

3. Deliver the intervention worksheets / homeworks. 1 per week and the sheets are numbered 1-6. Allow enough time (15 minutes) to discuss each homework task generally in the next lesson. You don't mark the work sheets but please take them in and keep them. In addition, it is assumed that as your pupils are approaching a high stakes' test, you are using practice questions / tests in the classroom. You must make the link between what you are doing in class (practising tests) and the cognition of the testing effect. My phrase is '*your testing routes are well oiled now*' in preparation for the tests. Ultimately, you are aiming for your pupils to feel self-efficacious (belief in their capability) at test taking.

4. After the high stakes' test, as soon as is physically possible, please administer the same two questionnaires (CTAS and HCAY) but they are marked post-test. Please would you let me know how soon after you were able to administer the post-test questionnaires.

5. As soon as possible after the high stakes' test, ask the pupils the pupil questions and please take notes of the responses.

6. Please complete the teacher questionnaire

7. Please send by registered post to: Helen Barsham, XXXX XXXXX or email securely and provide a password.

Checklist for inclusion:

Pre-test CTAS completed questionnaires (with pupil initials)

Pre-test self-efficacy questionnaires (with pupil initials)

Post-test CTAS questionnaire (with pupil initials)

Post-test self-efficacy questionnaire (with pupil initials)

All cognitive worksheets – (if you wish)

Pupil responses to questions post-test

Teacher questionnaire

General points:

All the data are anonymised in my analyses, but I need to be able to measure the effect of the intervention so please use pupil initials for pre and post CTAS and pre and post HCAY please).

Thus, for each pupil you will collate 4 questionnaires: 2 pre-test (CTAS and self-efficacy) and 2 post-test (CTAS and self-efficacy) and all four will be for the same pupil with their initials on so that I can measure for each pupil both pre and post-test.

And thank you!

Teacher Questionnaire:

1. Have the children understood *the testing effect*?

2. Do you think knowing about the testing effect has helped them in their tests?

3. What is more powerful, knowing about the testing effect or practising tests or a combination of both?

4. Which cognitive worksheets did you use?

5. Which ones were the best?

6. What did not work?

7. Did the Carol Dweck TED video work?

8. Does it belong in this intervention?

9. Any further advice, thoughts or tips

Questions for students (post-test)

Please would you record the responses –hand-written or typed notes are fine- thank you.

(Taken from the original pilot on 24.1.18 and conducted in the classroom)

Q1: How did you feel when you went into the test room?

Q2: How did you feel after answering the first few questions?

Q3: Were you able to use any of the strategies, we have been discussing? (wait for response)

Did you feel better? More confident? -

Q4: Did doing practice tests help?

Q5: Did learning about cognition (the testing effect) help and why?

Q6: Did you feel well prepared and why?

Cognition worksheets:

Week 1: Photocopy 2 pages onto A3 pp13-14

Week 2: Photocopy 2 pages onto A3 pp15-16

Week 3: Single sheet p17

Week 4: Single sheet p18

Week 5: Single sheet p19

Week 6: Single sheet p20

Cognition worksheet 1 (these x2 pages need to go as an A3 sheet)

Retention = what the memory can remember for a long time
Material = the topic being studied
Free recall = what can you remember without being prompted

Henry Roediger (Roddy to his friends) and Jeffrey Karpicke wrote many papers about cognition.

What do you think the following title really means?

'Test Enhanced Learning – Taking Memory Tests Improves Long Term Retention'

Roddy and Jeffrey said 'In two experiments, students studied prose (writing) passages and took one or three immediate free recall tests, without feedback, **or** they re-studied the prose material the same number of times as the students who received the tests. Students then took a final retention test at either 5 min, 2 days, or 1 week later. When the final test was given after 5 min, repeated studying improved recall compared to repeated testing. *However,* on the delayed tests, the earlier testing had produced greater retention than studying, even though repeated studying increased students' confidence in their ability to remember the material. Testing is a powerful means of improving learning, not just assessing it.'

What worked better? Studying or testing? Give reasons for your answer.

Do you like tests?

Roddy and Jeffrey also said:

'Practising the skills during learning that are needed for retrieval (testing) improves retention on memory tests. Although restudying the passages gives students all of the information again, testing allowed the students to practise the skill that they need for future tests and improved performance after a delay.

Can you summarise this in your own words?

Is this something you would like to do?

How might this help you?

Finally

‘Students may prefer repeated studying because it produces short-term benefits, and students often learn ineffective learning strategies because they base their predictions of future performance on what produces rapid short gains. Although students in the repeated study conditions predicted they would perform very well a week later (compared to the other students who were just being tested) - they actually performed the worst.

What do you think to Roddy and Jeffrey saying this?

Your final thoughts please – what have you learned?

Cognition worksheet 2: (these x2 pages photocopy onto A3)

New information coming…Interleaving = mixing up two or more topics when you study, so study a bit of Math and then some Geography - just mixing up what you do – not doing one subject all at once.

New information coming…Spacing = If you **interleave,** you will naturally space. So, imagine learning 20 French words then move onto your History revision. Hopefully you will forget some of the French words because letting the memory forget is a good thing. It means you need to make an effort to remember, and the information will stay in your memory for longer. Get it?

Last week - Remember Roddy and Jeffrey? What messages can you remember?

New information coming…I'd like to introduce Bob Bjork. He says we need '*desirable difficulties'* to learn well. These learning conditions can help you remember – even when you're old!

1. **Interleaving** instruction on separate topics rather than grouping by topic (called blocking)
2. **Spacing** rather than massing study sessions on a topic
3. **Using tests** rather than studying again (Henry and Jeffrey approve)
4. **Vary** the contexts - even **moving to a different room to study** can help

What is **interleaving**?

What is **spacing**?

What is better **testing** yourself or re-reading?

Is moving to another room to study a good thing?

Bob is a very wise man and has said many wise words about how best to learn. We would be very daft if we don't take his advice.

'Good test performance following an all-night cramming session is certainly rewarding, but little of what was recallable on the test will remain recallable over time. In contrast, a study schedule that spaces study sessions on a particular topic can produce both good exam performance and good long-term retention.'

What does Bob mean?

Roddy and Jeffrey ran a series of experiments in 2008. They found that ***"study, test, test, test"*** helps children learn more and perform better in tests than **'*study, study, study, test'***. Can you summarise this in your own words?

Cognition worksheet 3:

Can you still remember the main messages from Jeffrey and Roddy?

What about our new wise friend Bob? What were the four ***desirable difficulties*** – can you explain them?

Cognition worksheet 4:

Bob and Roddy and Jeffery had a meeting this week and wondered if they had got any messages across to you. They are so desperate to give you good learning skills. Bob is quite worried, so he wrote you a letter:

'To those of you, who are students, we hope we have convinced you to take a more active role in your learning by introducing ***the testing effect*** into your own study activities. Above all, try and rid yourself of the idea your memory works like a tape or video recorder and that re-reading the same material over and over again will somehow write it into your memory. Rather, assume that learning requires an active process of interpretation, that is mapping new things we are trying to learn on to what we already know.

Be aware, too, when re-reading a chapter or your notes, that prior exposure (having read them a few times before) creates a sense of familiarity that can easily be confused with understanding. And perhaps most importantly, keep in mind that retrieval (recalling the information), much more than restudying acts to modify (to change) your memory by making the information that you practise retrieving more likely to be recallable again and in different contexts. In short, try to spend less time on the input side and more time on the output side, such as summarizing what you have read from memory, or getting together with friends and asking each other questions. Any activities that involve testing yourself – that is activities that require you to retrieve or generate information, rather than just presenting information to yourself -will make your learning both more durable and flexible'.

Finally, we cannot overstate the importance of learning how to manage your own learning activities. In a world that is ever more complex and rapidly changing, and in which learning on one's own is becoming ever more important, learning how to learn is the ultimate survival tool'

Please reply to Bob Bjork (generally hangs out at the University of California, Foglab) and advise him how you are going to 'learn how to learn so that you have the ultimate survival tool' for the future.

I think you could probably say 'Dear Bob not Mr. Bjork' – I will find out his exact address and send the letters on…..

Cognition worksheet 5

It's not just the men who know how to learn – Carol Dweck has spent her whole life trying to understand how children think so that she can give you the best advice.

What are the key messages from the 'not yet' video?

Key messages: Scientifically, what is happening to the brain?

Can you recall a time when you used 'The Growth Mindset?'

How might this help you in evaluative situations like assessments? Driving tests? (It will happen one day!) Music exams? Anywhere else for 'not yet?'

Cognition worksheet 6:

A friend has texted you today. She is really worried about the exam tomorrow. She is worried she will fail and that she will disappoint her parents, her friends will laugh at her. She took an exam before and failed because although she had worked really hard, she became so nervous, she couldn't remember anything. She remembers how upset she was then.

You know she has been studying hard for the exam. She has been testing herself and looking up answers when she has forgotten. Two nights ago, all of you got together and tested each other. Your teacher regularly tests you in class and you have been following a cognition programme about *the testing effect.*

What advice will you give this friend?
NB: A list of salient references was also included for the schools delivering the intervention and omitted here due to space and all references are already included in the main References.

Appendix C

New Theory of Disuse

New Theory of Disuse – a Possible Further Intervention to Help Students Learn How to Learn

The New Theory of Disuse (Bjork & Bjork, 1992) has been included here for two reasons: Firstly, it may help to inform the creation of an intervention that uses cognition of memory. Secondly, I have created a highly illustrative example from the theory that may aid the reader in visualizing how retrieval induced forgetting works; this is a central tenet to understanding how the testing effect works.

I found it helpful to create a literary picture about what was being explained to aid my understanding and as such have created a mentally illustrative model that may help inform a future intervention for this area, where an understanding of how memory works is important for learning how to learn.

Bjork and Bjork (1992) explain the New Theory of Disuse. It is about two strengths of memory: the storage and retrieval strength. Increases in retrieval strength are negatively correlated with increases in storage strength. If retrieval strength is high, then storage strength is lower. If there is lower retrieval strength, then storage strength is higher. In other words, the more difficult something is to recall, the longer it will stay in memory, as long as it can be remembered in the first place. Conversely, if something is easy to remember, it may be forgotten quickly.

Storage strength is how well learned an item is and the degree of its strength is shown through the ease of its retrieval. To grow an item in storage strength, one needs to use opportunities for the study and recall of items. There is no limit on storage strength but there is a limit on retrieval strength. If retrieval strength is strong it is assumed to slow down storage strength's ability to store the knowledge. This is because the nature of retrieval strength is cue dependent and as certain items increase in retrieval strength, others are less recallable, but they are in the storage of the brain. An item's retrieval strength is larger, the more difficult the initial retrieval.

To use a visual analogy. Think of storage strength as a very large attic where knowledge is stored, and which can never be filled up. There is always plenty of space.

However, if you don't take an item out regularly to use, it becomes dusty and stuck at the back and can be difficult to find when you need it, but it is still there. The piles of knowledge at the entrance to the attic, which you use often aren't dusty and can be found quickly and are easily accessible, these are retrieval strength. Sometimes, if you're using a lot of recalled ideas from one recent pile, you may forget to take some of the items from another recent pile (retrieval induced forgetting). Ideally, you need to be able to reach the dusty old knowledge at the back and the way to do this is to keep moving the items in the attic around (interleaving, spacing and retrieval practice can do this) so that they all have a turn at being accessible and all have a turn at being at the back; better still if you can think of a different system for remembering where you have put them in the attic, like a map, or you have grouped some items together because they have a common theme (cues and schemas used for retrieval).

The New Theory of Disuse owes some of its features to Estes 1955b (Bjork & Bjork 1992). Memory has a semantic relationship to items' schemas and scripts already in the long-term memory. When memory encounters strong new information, there is plenty of storage 'in the attic'; remember there is unlimited storage space. It is also a magical attic where the space just keeps on increasing. The more knowledge we have, the more ways to store additional knowledge.

As humans, we are impressive at getting information into the brain (up into the attic), but not so good at getting it out (once it's got lost and dusty at the back). The retrieval process is fallible, erratic and cue (environmental, interpersonal, emotional, physical) dependent. Memory is a modifier (Bjork, 1975) because once we can recall where we stored the information (in the attic), we can find the same path again quickly. The information retrieved becomes more retrievable. "The testing routes are well-oiled". (Your attic map is working!) However, other information becomes less retrievable (a neglected pile in the attic or a lost route on the map). Some items become non recallable with disuse. These are the ones you cannot find in your attic at all, but they are still there.

Pan and Bjork (in press) suggest that a memory's current performance is its current retrieval strength. Storage strength cannot be observed and must be inferred (think of the attic space) but it is never lost. Successful retrieval has a larger impact on both storage strength and retrieval strength than re-study: the testing effect.

To make the most of memory skills, one should practise a delayed initial recall of the material and delay re-study. The key is to make the initial learning process difficult or effortful; allow some forgetting to have happened before retrieval is attempted. This method decreases retrieval strength and makes performance worse but enhances learning and later recall (e.g., Pyc & Rawson, 2009).

Appendix D

Teacher and Student Feedback

School Context	Teacher Comments
A, *n* = 15 Co-ed	Worksheet 3 was missed out. Students needed support with the language about *desirable* difficulties in week 2. Teacher felt that a video format would be better as this is the students' preferred learning style. The Carol Dweck video should stay in the intervention. Most students had understood *the testing effect* according to teacher feedback. 'A combination of practice tests and knowing about *the testing effect* worked.'
B, *n* = 4 Co-ed	All worksheets completed. Students liked week 5, the Dweck video and both letter activities. Students' comments were very positive; they believed learning about *the testing effect* had helped them when taking high stakes' tests. The teacher said 'A combination of using both *the testing effect* and practice tests supports the 'why''
C, *n* = 15 Girls	Worksheets 5 and 6 missed. Students liked the letter exercise (week 4) and needed support with the language of *desirable difficulties.* The teacher felt that the intervention 'would suit older students.' Student responses were all completed individually in the feedback and 9 out of 15 students said that they felt the intervention had helped them in taking tests.
D, *n* = 19 Co-ed	All worksheets completed. Some help needed with explaining the language of *desirable difficulties* at first. The Carol Dweck video should stay in the intervention. Students had enjoyed the worksheets, especially writing the letters and had understood *the testing effect* and felt it had helped them in their tests. The teacher said, 'I'd like to run this again'.
E, *n* = 20 Girls	All worksheets were completed. Students liked the letter in week 4 and the advice letter week. Students did need help with the language about the wider *desirable difficulties*. Students understood *the testing effect* and felt that it helped them in their tests. The teacher said 'Yes, they understood this, it worked well'. (Although, I delivered the intervention, I did not take the student feedback. This task was completed by the students' form teacher.)
F *n* = 12 Co-ed	All worksheets were completed. Students liked both letter activities. Students needed help with the *desirable difficulties'* terms in week 2. Students understood *the testing effect* and felt that it helped them in their tests.

Examples of Student Responses to Worksheets and Pupil Questions from Participating Schools

‘I think they (Jeffrey and Henry) are saying that most people feel more comfortable with studying a lot, but it is better to study a bit and then test yourself actually’.

‘I have learnt that testing repeatedly improved greater retention for the actual test (exam) and that repeated studying was better for short gains but not long weeks of having to remember’

‘Learning about cognition helped because otherwise I would have kept studying rather than testing’

‘Learning about learning to learn helped me realise how we learn’

‘Learning about study, test, test, test worked. I did not know it before, but it worked when I tried it’

‘Forgetting things isn’t bad because when you forget something there is more space in your mind to learn new stuff so it’s ok to forget.’

‘Knowing about the testing effect helped, it’s proved so I know it actually works’

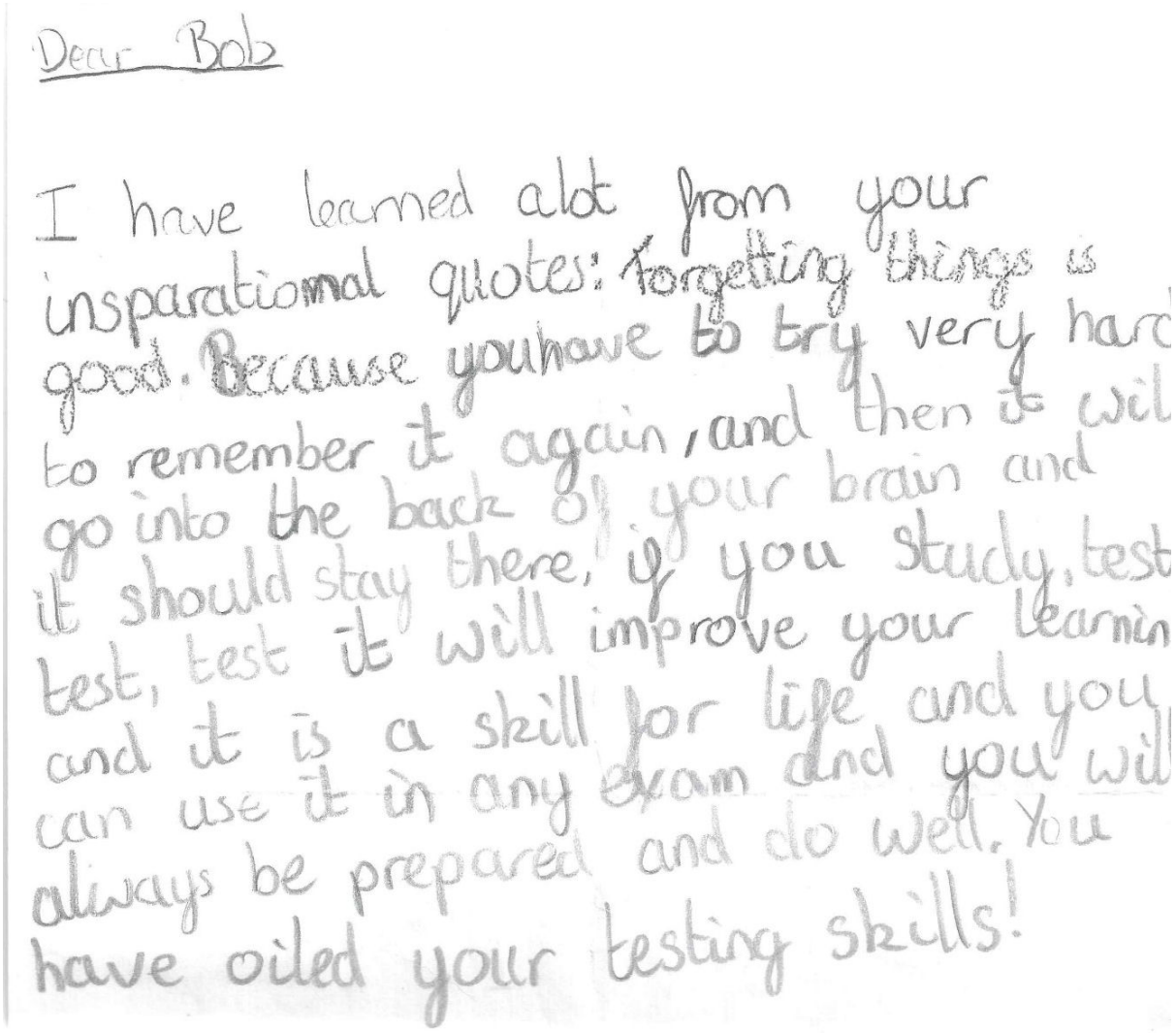
‘Learning about the testing effect made me feel more confident’

‘Putting a lot into your brain helps short term memory but study, test, test, test, helps you remember it for longer.’

‘I think testing works better because under pressure, it is more likely to stay in your head’

‘Testing works for me because it keeps getting hammered into my brain’

Letters to Bob



Dear Bob

I have learned alot from your insparational quotes: forgetting things is good. Because youhave to try very hard to remember it again, and then it will go into the back of your brain and it should stay there, if you study, test test, test it will improve your learning and it is a skill for life and you can use it in any exam and you will always be prepared and do well. You have oiled your testing skills!

DEAR: Bob

Hello! I have learned a lot of things from you e.g. Forgetting things is good for your brain. Also you told me that if I cram for exams then I'm more likely to forget things in the long term, but if I revise over a long period of time then I more likely to remember things.

Dear Robert,

I am writing to you to explain how I am going to show metacognition (learning about learning) in everyday life. Recalling is when you have studied material or done something before and you have been tested or asked to remember it. It will help people for when it comes to an exam, because Study, study, study test is much better for short-term retention, e.g. you forgot about an exam/test you had the next day and you just revised the same thing over and over again it would be alright if it was the next day. But practising long-term retention you would have to use the method of study test, study test and interleaving, which is where you mix up the subjects e.g. learning about victorians and cavemen at the same time.

Your information has given me alot of ideas and thoughts on metacognition. This will really help me will anxiety when it comes to tests. Now I feel prepared for any test as metacognition will help me see what I really need to do. If I come to a situation where something has happened and I can't think, I will narrow my mind down to focus using metacognition.
Thank you for your research and information.
Yours sincerely,

Dear Robert,

A way I will use metacognition is, when I am learning, for say that Latin test we had the other day, I can learn the contence in ~~diffrent~~ rooms and in distre positions. So It could learn outside upside down or the living room with my cat. What I ~~f~~ hope you understand is that I find metacognition a hard th to learn about. I have a small case of dislexea so I feel that gives me disegvanteges.

Even though I have disedvantges I understand how study ~~th~~ test test helps improve your long turm memor and keeps information in for a longer period of tim I have a sister and she is called . My mother he her with some metacognition learning. So, what my sister did was she combined her geography form last ye with her biolagy. I belive this is a very good way to practise the use of metacognition.

Your Sincirly,

Dear Bob,

You and your friends have taught me quite alot so far from the metacognition papers and I would just like to say thankyou.

I will now tell what I have learnt and why. It is a hole lot better to study, test, test, test than doin study, study, study, test because if you get a question wrong you can remember how for next time but if you only do the test once and get a question wrong your brain will forget it and you will never learn. Im going to learn how to learn by studying a bit on lots of subjects then testing myself more on various subjects. I will also remember to have more convidence in myself.

Dear Robert,

I am writing to tell you not to be worried because we have learnt so much from you, Henry and Jeffery. We have been using your desirable difficulties: Interleaving instruction on Separate topics rather than grouping by topic, Spacing out your studying instead of doing it all at once, test rather than studying and moving to a different room. You have also told us that forgetting helps strengthen your brain cells so you can remember new information and skills, and when a skill can be retrieved easily from memory little or no learning results

Dear Robert Bjork,

I have learnt that forgetting things is good, because pulling things from your brain is better than being able to remember things easily, because it's able to remember things easily because it's an event happening and your brain will remember it.

It's better to study test test test than study study study test because you will get used to the test and will have practice

Advice Letters to Friends

What advice will you give this friend?

Dear

Don't panicke youll be fine. It scientificly proven that if your positive and not panicked you will do fine. Also study, test, test emprovs long turm memory and thats what youve done. Don't worry you'll do great.

Love from

What advice will you give this friend?

I would tell her not to think about the past to just focus on the test also to think about to just keep following the 'not yet' and ~~stude~~ 'study, test, test, test' and also tell her she will do fine

Also her friends won't laugh and if you do your best then your parents won't care.

Just think if you have forgotton something make your brain hurt and if you can't get it think 'not yet' but if you keep testing yourself just imagine you are testing yourself and you will do fine.

What advice will you give this friend?

I would tell her to stop worrying because her to reassure her and tell her that she has been following the 'best method' that there is proven by science and testing and that she has done the best she could do and she couldn't do done any more so she should get a good nights rest and to believe in herself.

What advice will you give this friend?

The advice I would give to this friend is that she should stop worrying so much and instead learn ~~about~~ from the mistakes that she did. SSST doesn't always work as she has done, because it means that you don't know what to expect in the exam but if she tests herself she will have an idea of what she has to do and learn it.

When learning is hard, it is stronger and lasts longer!

The friend should learn this phrase as it will help her remember that testing is good.

Appendix E

Ethics

Sample Parental Permissions Letter for Schools

(the letter was adapted for the schools in the control group)

Dear Parents,

As you are aware, (name of school) is wholly committed to caring for and promoting the wellbeing of your sons and daughters.

We have an opportunity to work with a research degree student from the University of Cambridge. She is also headteacher.

The research is aimed at increasing self-efficacy in 10 -to11- year- olds in evaluative situations. Essentially, a six-week homework intervention teaches the children about the cognition of the 'testing effect' and a pilot study showed that the pupils felt more confident in an examination situation as a direct result of learning about how memory works for test situations.

We need your permission for your son/daughter to complete 2 short questionnaires about attitudes to test and feelings of self-efficacy. These are completed both pre and post-tests, so a total of 4 questionnaires. In addition, the children will be asked their opinions about the intervention and we need your permission to be able to use their comments.

All data are anonymised and will be stored electronically or on paper and may contribute to scientific papers and presentations.

No individuals are named, and the research has been reviewed for its ethics, by the University of Cambridge.

You are free to withdraw your consent / permission at any point, simply by emailing the address below.

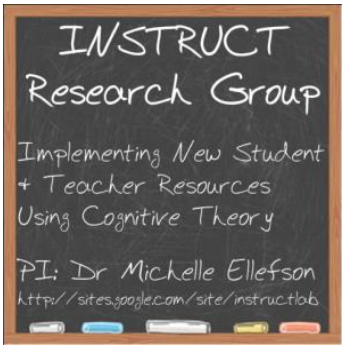
Please do feel free to contact Helen Barsham for further information and her details are included below. In addition, we attach a summary of her research so far.

Please email ----------with your permission.

Yours Sincerely,

Helen Barsham –doctoral student at:

UNIVERSITY OF CAMBRIDGE
Faculty of Education



hb479@cam.ac.uk

INSTRUCT Research Group Psychology and Education, Web: *http://sites.google.com/site/instructlab/*

RESEARCH ETHICS REVIEW CHECKLIST FOR FACULTY OF EDUCATION
(*electronic version of the original)*

Section A: Details of the Project

Student Name	Helen Barsham
Email	hb479@cam.ac.uk
Supervisor	Michelle Ellefson
Supervisor email	mre33@cam.ac.uk
Registration Report Title	Can an Intervention Programme, Which Uses Metacognition and Practice of the Theory of the Desirable Difficulty of the testing effect Reduce Test Anxiety, in 10-to11- year-olds*, In Evaluative Situations?

Section B: Checklist

Code of Practice relating to Educational Research		
1a	Have you read the *Revised Ethical Guidelines for Educational Research* (2011) of the British Educational Research Association (BERA)? (if you have not read it, the latest version is available at http://www.bera.ac.uk/researchers-resources/publications/bera-ethical-guidelines-for-educational-research-2011	Yes
1b	Is this Code relevant to the conduct of your research? If you have answered 'no', please briefly explain why:	Yes
1c	Do you agree to subscribe to the Code in carrying out your own research?	Yes
2	Are there any aspects of your proposed research which, in the context of BERA's Code of Practice, might give rise to concern amongst other educational researchers?	No
If you have answered 'yes', please briefly list possible causes for concern below:		
a		
b		
c		
3a	Will you be analysing an existing data set that has already been collected by someone else?	No

3b	If you answered YES: can you confirm that the data you will be using is *either* Already available in the public domain for anyone to analyse Or You have been given permission by the owner of the data set to undertake your own analysis and results[1]	Yes/No
4	Will you be collecting your own research data for the study (through such techniques as interviewing people, observing situations, issuing questionnaires etc)? ***nb. If you have answered NO to this question, you may proceed to Section C and need not answer any further questions in this section.***	Yes
Obtaining 'Informed Consent'		
5	Are you familiar with the concept of 'informed consent'? (if you are not familiar with this concept you should first consult the following source: page 5 of the BERA guidelines above).	Yes
6	Does your research involve securing participation from children, young people or adults where the concept of 'informed consent' might apply? *Permission is likely to be needed to report any information about people or institutions that is not in the public domain, and which you have been able to obtain due to your privileged access to the research site(s) in whatever capacity* [2]	Yes
If you have answered 'yes' to Question 6 above, please answer the following questions.		
7a	Do you believe that you are adopting suitable safeguards with respect to obtaining 'informed consent' from participants in your research in line with the Code of Practice?	Yes
7b	Will all the information about individuals and institutions be treated on an 'in confidence' basis at all stages of your research including writing up and publication?	Yes

[1] *this permission should only be given if the owner of the data can make it available for secondary analysis on the basis of the informed consent they obtained from their original participants*

[2] Professional work (such as teaching) can involve the collection of evidence to better understand problems/issues and to evaluate innovative practice - leaving practitioners with the question of when these activities become formal research requiring informed consent. This comment is meant to highlight how the collection of data for public reporting beyond the institution (e.g. in a thesis) should be considered as a key criterion for deciding when informed consent is required.

7c(i)	Will all the information collected about the institution(s) where research is based be presented in ways that guarantee the institution(s) cannot be identified from information provided in the report? *Note: in a thesis written by a researcher about a research context where they have a publicly acknowledged role, it is difficult to disguise the identity of the institution whilst also providing the expected detail of the researcher's relationship with the research context.*[3]	No
7c(ii)	If not, has the appropriate responsible person given approval for the research on the understanding that the identity of the institution cannot be protected in the report of the research?	Yes
7c(iii)	Will all the information collected about individuals be presented in ways that guarantee their anonymity? *Note: a person with a named role, or having a specific set of reported characteristics that is unique in the research context, cannot be assured of the anonymity when the identity of the research site cannot be protected.*	Yes
7c(iv)	If not, have these issues been explained to the relevant participants (and appropriate gatekeepers in the case of children or other vulnerable participants)?	Yes/No
The Involvement of Adults in the Research		
8a	Will your research involve adults?	Yes
If you have answered 'yes' to Question 8a above, please answer the following questions; otherwise move to Question 9.		
8b	Will these adults be provided with sufficient information *prior* to agreeing to participate in your research to enable them to exercise 'informed consent'?	Yes
8c	Will the adults involved in your research be in a position to give 'informed consent' themselves with respect to their participation?	Yes

[3] At present the implicit assumption is that anonymity is always desirable*, and is always achievable. In many studies these assumptions are sound. However, a practitioner (e.g. teacher) reporting research into their own practice/institution in a thesis would normally need to be explicit about their professional relationship to the research context to give an authentic account of their research. As the staff lists of many educational institutions are in the public domain and often readily found by a web search, a thesis by a named member of staff allows the institution to be readily identified from the name of the thesis author.
Given that an institution can readily be identified, this also has consequences for the degree of anonymity that can be promised to participants - for example those with named roles such as Head of Year 11, Student Voice Coordinator, Head Prefect, etc, or those identifiable from detailed reported characteristics.
* Some institutions or participants may welcome being acknowledged by name in a thesis, and their views should be taken into account and balanced against other considerations.

8d	Will these adults be able to opt out of your research in its entirety if they wish to do so by, for example, declining to be interviewed or refusing to answer a questionnaire?	Yes
8e	Will these adults be able to opt out of parts of your research by, for example, declining to participate in certain activities or answer particular questions?	Yes
The Involvement of Children, Young People and other potentially Vulnerable Persons in the Research		
9a	Will your research involve children, young people or other potentially vulnerable persons (such as those with learning disabilities or your own students).	Yes
If you have answered 'yes' to Question 9a above, please answer the following questions; otherwise move to Question 10. In educational and social research 'informed consent' regarding access is often given by a 'gatekeeper' on behalf of a wider group of persons (e.g. a head or class teacher with respect to their pupils, a youth worker working with young people, another person in an 'authority' position).		
9b	Who will act as the 'gatekeeper(s)' in your research? Please list their position(s) briefly below and, where this is not self-evident, describe the nature of their relationship with those on whose behalves they are giving 'informed consent'. **The researcher cannot act as the gatekeeper (see 9g below)**	
i	headteacher Sept 2018 –Jan 2019	
ii	Regional Principal Jan 2019 ff	
iii		
9c	Will you be briefing your 'gatekeeper(s)' about the nature of the questions or activities you will be undertaking with the children, young people or other potentially vulnerable persons involved in your research?	Yes
9d	If another person (such as a teacher or parent of a child in your study) expressed concerns about any of the questions or activities involved in your research, would your 'gatekeeper(s)' have sufficient information to provide a brief justification for having given 'informed consent'?	Yes
9e	If unforeseen problems were to arise during the course of the research, would your 'gatekeeper(s)' be able to contact you at relatively short notice to seek advice, if they needed to do so?	Yes
9f	Could your 'gatekeeper(s)' withdraw consent during the research if, for whatever reason, they felt this to be necessary?	Yes
9g(i)	Are you undertaking research into your own professional context/institution (e.g. with students in a school where you work)? If you answered 'Yes' then you should identify (in 9b above) a suitable senior person who has agreed to act as an independent point of contact for participants to act as the gatekeeper, and answer the following two questions:	Yes

9g(ii)	Will you ensure that other people in the research context are aware of the identity of the gatekeeper?	Yes
9g(iii)	Will you take reasonable precautions to ensure that research participants (and where appropriate their parents/guardians) know that they should contact the gatekeeper (and not you) if they have any concerns about the research?	Yes
Other Ethical Aspects of the Research		
10	Will it be necessary for participants to take part in the study without their knowledge and consent at the time? (eg covert observation of people in public places)	No
11	Will the research involve the discussion of topics which some people may deem to be 'sensitive'? (e.g. sexual activity, drug use, certain matters relating to political attitudes or religious beliefs)	Yes
12	Does the research involve any questions or activities, which might be considered inappropriate in an educational setting?	No
13	Are drugs, placebos or other substances (e.g. food substances, vitamins) to be administered to study participants or will the study involve invasive, intrusive or potentially harmful procedures of any kind? *If you have ticked 'Yes' it is vital to refer the matter to the Faculty Research Office for onward reference to the University Insurance Section.*	No
14	Will blood, tissue or other samples be taken from the bodies of participants?	No
15	Is pain or more than mild discomfort likely to result from the study?	No
16	Could the research involve psychological stress or anxiety or cause harm or negative consequences beyond the risks encountered in normal life?	Yes
17	Are there any other aspects of the research which could be interpreted as infringing the norms and expectations of behaviour prevailing in educational settings?	No
18	Are there any other aspects of the research which could be to the participants' detriment?	No
19	Will the study involve prolonged or repetitive testing?	No
20	Will financial inducements (other than reasonable expenses or compensation for time) be offered to participants?	No

SECTION C: Interpretation of Results

If any of your answers coincide with the response options having a coloured background, then you should assume that further discussion involving Stage 2 procedures is required because some aspect of your proposed research is likely to be 'ethically sensitive'. In practice, many issues can be resolved at this stage. In practice, many issues can be resolved at this stage.

Members of staff should be especially careful about research involving their own students (question 9g).

If you have ticked 'yes' in response to one or more of questions 10 to 20, both Stage 2 ***and*** *Stage 3 clearance will definitely be required.*

Stage 2 Clearance
Any 'ethically sensitive' responses identified above should be discussed with a 'knowledgeable person of standing'. In the case of students within the Faculty, this person will, in almost every case, be the person supervising your research.

On completion of the discussion, the 'knowledgeable person of standing' is asked to choose one of the following three responses, to delete the other two and to affirm their views by adding their signature.	
a	I have discussed the ethical dimensions of this research and, as outlined to me, I do not foresee any ethical issues arising which require further clearance.
b	~~There may be some ethical issues arising from this research. I think it would be prudent for the researcher to seek further advice and, possibly, Stage 3 clearance.~~
c	~~Ethical issues arise in this research, which require further discussion; my advice is that Stage 3 ethical clearance should be sought.~~

Supervisor Name/ Signature	Dr Michelle Ellefson
Date	August 2018

* This refers to Year 6 in the primary phase of education.

7ci – I have not mentioned the school where I work by name. I have referred to the type of school it is for to explain aspects of my study. All stakeholders, in all linked institutions, will never be mentioned by name nor their schools identified. I will only ever refer to the type of school.
All parties will always be aware that research is being carried out.
No specific classes or categories of teacher will be mentioned and I will seek out the most general label at all times E.g. 10 -11 year olds.

8a – I have not used adults in the pilot study but hope to use teacher questionnaires in the main study.

11 – I will be asking pupils about their test anxiety. However, the choice of instrument is designed for this age-group specifically and will be administered at least 6 weeks ahead of any test to avoid priming. It is unlikely that a child will become distressed at participating in this research but as all staff who, have contact with a child will be aware of this research, a buddy can be established for the child - someone they like to talk to. There will also be form teachers available and parents will be contacted if a child becomes unhappy. Pastoral and PHSE time, counsellors can be made available too.

16 – It is important to do this research because the purpose is to enable children to feel confident in testing situations. The choice of measurement has been chosen carefully and piloted with children. In the main study, the use of a self-efficacy questionnaire will also be trialled with children before it becomes embedded. Any child who becomes distressed will have access to the usual pastoral care systems in a school and be supported through time, conversation, parental contact and support, PHSE and counselling.

NB*: This is an electronic copy of the original form that is held in the higher degrees' office.*

Appendix F

Results from the Pilot

Table F.1

Pilot - Paired t Test Results

Question	*M*	*SD*	*SE*	*t*	*p*	Cohen's *d*
Thoughts						
Will I pass	1.00	1.18	0.36	2.80	.02	0.85
Get bad result	0.36	1.21	0.36	1.00	.34	0.30
Hard remember answers	0.36	0.50	0.15	2.39	.04	0.72
Worry about failing	0.36	1.12	0.34	1.08	.31	0.32
Do something wrong	0.64	1.12	0.34	1.88	.09	0.57
What results will be	0.36	1.29	0.39	0.94	.37	0.27
Answers are right	0.45	0.93	0.28	1.61	.14	0.48
Have studied more	0.82	0.98	0.30	2.76	.02	0.82
Most answers wrong	0.64	1.03	0.31	2.06	.07	0.62
Worry how hard	0.73	1.19	0.36	2.03	.07	0.61
What if I fail	0.36	1.03	0.31	1.17	.26	0.35
How badly I'm doing	0.36	1.03	0.31	1.17	.27	0.35
Worry parents say	−0.09	0.54	0.16	−0.56	.59	0.16
Off Task Behaviour						
Look around room	0.27	1.10	0.33	0.82	.43	0.25
Play pen pencil	−0.36	1.36	0.41	−0.89	.40	0.26
Check the time	0.55	0.69	0.21	2.63	.02	0.78
Hard sit still	0.18	1.25	0.40	0.48	.64	0.14
Look at other people	0.82	1.08	0.33	2.52	.03	0.76
Try finish quickly	−0.73	1.27	0.38	−1.90	.08	0.57
Tap my feet	0.27	1.35	0.41	0.67	.52	0.20
I stare	0.45	1.37	0.41	1.10	.30	0.33
Autonomic Reactions						
Heart beats fast	1.00	0.77	0.23	4.28	.00	1.30
Feel nervous	0.82	0.75	0.23	3.61	.00	1.10
Face feels hot	−0.09	1.22	0.37	−0.25	.81	0.07
Belly feels funny	0.73	1.01	0.30	2.39	.04	0.72
Head hurts	0.27	0.9	0.27	1.00	.34	0.30
Feel warm	0.09	1.38	0.41	0.22	.83	0.06
Hand shakes	0.27	1.01	0.30	0.90	.39	0.27
Go bathroom	−0.09	1.14	0.34	−0.27	.59	0.07
Feel scared	0.73	1.01	0.30	2.39	.04	0.72

Note. Variables in Pre and Post-test Questionnaires in the Sample; Arranged in Subscales. *N* =11. *Degrees of Freedom* for all of these results are 10.

Student Responses from the Pilot Study

‘I think they (Jeffrey and Henry) are saying that most people feel more comfortable with studying a lot, but it is better to study a bit and then test yourself actually’.

‘I think the testing worked better because if you study once and then test yourself three times.’

‘I have learnt that testing repeatedly improved greater retention for the actual test (exam) and that repeated studying was better for short gains but not long weeks of having to remember’

‘When you study loads and loads for an exam, when the exam is done you forget about it but if you study and then test you will remember.’

‘Robert is saying roughly the same as Henry and Jeffrey, however, he also states that it is important to forget and make the brain work harder to remember. The harder you remember the better the long-term memory.’

‘Forgetting things isn’t bad because when you forget something there is more space in your mind to learn new stuff so it’s ok to forget.’

‘Forgetting is like a mistake you learn from it. When you learn something test yourself, so you learn for life, it is useless to learn something then forget like a doctor test it is important to remember for future reference.’

Letters to Bob from the Pilot Study

Dear Robert,

We are going to 'learn how to learn' at school by learning about metacognition and reading and writing about how remembering works and how to remember things. Also, for example, when we do our spelling tests, we are practicing to not cram it in all at the last possible moment or do it every day for a whole week. We get a sheet and we do it day 1, day 3, day 5 so we spread it out. This forces you to remember. Then, your brain, while you are remembering is putting that information into your deeper memory. In spellings we get tested on that week's spellings and a few of the week's before. It is also better to study, test, test, test: but it is not good for retention to study, study, study, test.

We, as students, usually prefer studying to testing because some people get nervous or feel pressure, but testing makes you think more which means (as Mrs. Barsham told us we can) grow your brain! Growing your brain means you have more brain cells, which expands your knowledge of subjects. As someone quoted about Einstein 'He knew a little about everything'.

If you also mix topics up, for example, writing a script in French improves your script writing and your French. Also changing the setting for example learning Latin in one room one day and a different room the next will help you. I hope you are well.

Your Sincerely,

Dear Robert,

I have tried doing your technique for one of my exams and it worked because I got full marks. So, thank you and I find using your technique really helps me to understand what I'm doing, and it goes into my long-term memory as well. It has really helped me. I have learned how to deal with tests, so I don't get worried about them anymore. In fact, I feel confident when I go into them. I got my parents to test me each night and morning instead of just staring at words when I revise and hoping to remember them. I feel like I will remember that test forever, I think also that we should learn about one thing say food in History and do that subject in lots of different historic stages. You have really helped me.

Yours Sincerely,

Dear Robert,

I read your letter and I have decided that when I study for future tests, I will test myself more than reading my notes over and over again because once I understood how my brain works, I can understand how to change the way I feel about a test or tests. Once, I know how my brain works, I can learn the difference between being familiar with my notes and remembering and understanding it even if it's just getting together with friends and testing each other on notes and questions, I know that I am more likely to remember it. Therefore, I wouldn't have to worry about the test, and I think more pupils should be given the opportunity to learn this way so that they have a better future.

Appendix G

Results of the Children's Test Anxiety Scale

Descriptive Statistics

Table G.1

Means, Standard Deviations, Skew and Kurtosis, Pre-test, Intervention Group (n = 85)

Question	*M*	*SD*	Skew	Kurtosis	*SE*
Q1 Wonder if pass	2.35	0.80	0.13	−0.48	0.09
Q2 Heart beats fast	2.21	1.00	0.28	−1.07	0.11
Q3 Look around room	2.36	0.91	0.54	−0.58	0.10
Q4 Feel nervous	2.74	1.03	−0.20	−1.17	0.11
Q5 Get a bad result	2.06	0.90	0.46	−0.66	0.10
Q6 Hard to remember	2.13	0.84	0.35	−0.52	0.09
Q7 Play pen /pencil	2.00	1.05	0.68	−0.80	0.11
Q8 Face feels hot	1.69	0.90	1.01	−0.10	0.10
Q9 Worry fail	2.16	0.84	0.40	−0.42	0.09
Q10 Belly funny	1.71	0.96	1.08	−0.08	0.10
Q11 Worry do wrong	2.59	0.94	0.30	−1.07	0.10
Q12 Check the time	3.00	1.05	−0.62	−0.92	0.11
Q13 What result will be	2.45	1.02	0.04	−1.14	0.11
Q14 Hard sit still	1.79	1.01	0.97	−0.37	0.11
Q15 Answers right	2.76	0.85	−0.22	−0.65	0.09
Q16 Studied more	2.01	0.96	0.70	−0.44	0.10
Q17 Head hurts	1.62	0.95	1.29	0.39	0.10
Q18 Look at others	1.84	0.90	0.81	−0.25	0.10
Q19 Answers wrong	1.93	0.88	0.75	−0.13	0.10
Q20 Feel warm	2.01	1.05	0.64	−0.87	0.11
Q21 How hard test	2.34	0.93	0.33	−0.17	0.10
Q22 Finish quickly	2.16	0.94	0.36	−0.81	0.10
Q23 Hand shakes	1.59	0.94	1.31	0.34	0.10
Q24 Happen if fail	1.95	0.97	0.70	−0.58	0.11
Q25 Go to bathroom	1.51	0.84	1.47	1.03	0.09
Q26 Tap my feet	2.28	1.12	0.34	−1.27	0.12
Q27 How badly	1.79	0.77	0.83	0.43	0.08
Q28 Feel scared	2.07	1.08	0.59	−0.97	0.12
Q29 Parents say	1.81	1.12	0.98	−0.61	0.12
Q30 I stare	1.92	1.00	0.86	−0.37	0.11

Table G.2

Means, Standard Deviations, Trimmed Mean, Skew and Kurtosis, Post-test, Intervention Group (n = 85)

Question	*M*	*SD*	Skew	Kurtosis	*SE*
Q1 Wonder if pass	2.46	0.85	0.35	−0.60	0.09
Q2 Heart beats fast	1.75	0.87	1.02	0.31	0.09
Q3 Look around room	2.42	0.98	0.36	−0.95	0.11
Q4 Feel nervous	2.27	0.96	0.42	−0.76	0.10
Q5 Get a bad result	1.82	0.86	0.78	−0.20	0.09
Q6 Hard to remember	1.86	0.79	0.54	−0.44	0.09
Q7 Play pen /pencil	2.04	1.09	0.54	−1.12	0.12
Q8 Face feels hot	1.51	0.81	1.64	2.02	0.09
Q9 Worry fail	1.85	0.88	0.81	−0.12	0.10
Q10 Belly funny	1.59	0.85	1.34	0.93	0.09
Q11 Worry do wrong	2.28	0.89	0.41	−0.56	0.10
Q12 Check the time	3.16	0.95	−0.82	−0.44	0.10
Q13 What result will be	2.47	1.03	0.14	−1.16	0.11
Q14 Hard sit still	1.85	0.98	0.97	−0.12	0.11
Q15 Answers right	2.52	0.91	−0.05	−0.83	0.10
Q16 Studied more	1.84	0.94	1.01	0.16	0.10
Q17 Head hurts	1.45	0.85	1.99	3.03	0.09
Q18 Look at others	2.02	1.10	0.75	−0.81	0.12
Q19 Answers wrong	1.68	0.88	1.18	0.57	0.09
Q20 Feel warm	1.85	0.96	0.95	−0.10	0.10
Q21 How hard test	2.18	0.97	0.43	−0.80	0.10
Q22 Finish quickly	2.11	0.91	0.54	−0.50	0.10
Q23 Hand shakes	1.40	0.76	1.95	3.14	0.08
Q24 Happen if fail	1.93	1.04	0.76	−0.72	0.11
Q25 Go to bathroom	1.27	0.62	2.64	7.27	0.07
Q26 Tap my feet	2.19	1.06	0.45	−1.06	0.12
Q27 How badly	1.60	0.80	1.37	1.45	0.09
Q28 Feel scared	1.73	0.88	0.86	−0.40	0.10
Q29 Parents say	1.81	0.99	1.02	−0.10	0.11
Q30 I stare	1.85	0.93	0.91	−0.09	0.10

Table G.3

Means, Standard Deviations, Trimmed Mean, Skew and Kurtosis, Pre-test, Control Group (n = 85)

Question	*M*	*SD*	Skew	Kurtosis	*SE*
Q1 Wonder if pass	2.41	0.93	0.17	−0.86	0.10
Q2 Heart beats fast	2.01	0.96	0.54	−0.77	0.10
Q3 Look around room	2.16	1.00	0.45	−0.88	0.11
Q4 Feel nervous	2.49	0.98	0.31	−1.04	0.11
Q5 Get a bad result	2.01	0.92	0.89	0.10	0.10
Q6 Hard to remember	2.13	1.02	0.54	−0.85	0.11
Q7 Play pen /pencil	1.79	1.10	1.10	−0.29	0.12
Q8 Face feels hot	1.72	0.96	1.22	0.42	0.10
Q9 Worry fail	2.22	1.07	0.47	−1.05	0.12
Q10 Belly funny	1.56	0.85	1.40	1.05	0.09
Q11 Worry do wrong	2.41	0.95	0.25	−0.90	0.10
Q12 Check the time	2.81	1.16	−0.27	−1.48	0.13
Q13 What result will be	2.55	1.10	0.03	−1.35	0.12
Q14 Hard sit still	1.94	1.09	0.76	−0.85	0.12
Q15 Answers right	2.88	0.99	−0.27	−1.18	0.11
Q16 Studied more	2.34	1.04	0.42	−1.03	0.11
Q17 Head hurts	1.49	0.81	1.67	2.12	0.09
Q18 Look at others	1.79	0.91	1.07	0.33	0.10
Q19 Answers wrong	2.18	1.06	0.54	−0.94	0.11
Q20 Feel warm	1.92	1.05	0.83	−0.59	0.11
Q21 How hard test	2.33	0.96	0.45	−0.76	0.10
Q22 Finish quickly	2.13	1.08	0.59	−0.94	0.12
Q23 Hand shakes	1.64	0.99	1.36	0.53	0.11
Q24 Happen if fail	2.07	1.13	0.64	−1.03	0.12
Q25 Go to bathroom	1.26	0.52	1.82	2.46	0.06
Q26 Tap my feet	1.98	1.08	0.72	−0.85	0.12
Q27 How badly	1.95	1.02	0.89	−0.35	0.11
Q28 Feel scared	1.86	0.94	0.87	−0.20	0.10
Q29 Parents say	2.13	1.11	0.57	−1.05	0.12
Q30 I stare	2.08	1.17	0.64	−1.11	0.13

Table G.4

Means, Standard Deviations, Skew and Kurtosis, Post-test, Control Group (n = 85)

Question	*M*	*SD*	Skew	Kurtosis	*SE*
Q1 Wonder if pass	2.39	0.96	0.20	−0.95	0.10
Q2 Heart beats fast	2.09	0.98	0.63	−0.60	0.11
Q3 Look around room	2.28	1.01	0.39	−0.94	0.11
Q4 Feel nervous	2.42	1.05	0.26	−1.17	0.11
Q5 Get a bad result	2.04	0.92	0.84	0.02	0.10
Q6 Hard to remember	2.09	0.88	0.64	−0.18	0.10
Q7 Play pen /pencil	1.89	1.08	0.88	−0.61	0.12
Q8 Face feels hot	1.68	0.99	1.24	0.25	0.11
Q9 Worry fail	1.91	0.91	0.84	−0.05	0.10
Q10 Belly funny	1.52	0.85	1.85	1.85	0.09
Q11 Worry do wrong	2.41	0.92	0.30	−0.77	0.10
Q12 Check the time	2.95	1.09	−0.56	−1.08	0.12
Q13 What result will be	2.46	1.05	0.17	−1.21	0.11
Q14 Hard sit still	2.13	1.14	0.51	−1.21	0.12
Q15 Answers right	2.55	1.04	−0.05	−1.20	0.11
Q16 Studied more	1.95	0.92	0.72	−0.36	0.10
Q17 Head hurts	1.38	0.71	1.74	2.01	0.08
Q18 Look at others	1.89	0.89	0.81	−0.05	0.10
Q19 Answers wrong	1.92	0.97	0.79	−0.41	0.10
Q20 Feel warm	1.87	0.90	0.84	−0.08	0.10
Q21 How hard test	2.31	1.00	0.36	−0.95	0.11
Q22 Finish quickly	2.11	1.00	0.57	−0.76	0.11
Q23 Hand shakes	1.59	0.97	1.51	1.00	0.10
Q24 Happen if fail	1.98	1.09	0.70	−0.92	0.12
Q25 Go to bathroom	1.33	0.78	2.52	5.54	0.08
Q26 Tap my feet	1.89	1.07	0.84	−0.66	0.12
Q27 How badly	1.75	0.87	1.02	0.31	0.09
Q28 Feel scared	1.75	0.95	1.16	0.36	0.10
Q29 Parents say	2.06	1.13	0.58	−1.13	0.12
Q30 I stare	1.84	1.04	1.01	−0.28	0.11

Table G.5

Mean, Skew and Kurtosis for the Original Thoughts' Subscale, Pre-test, Intervention Group (n = 85)

Question	*M*	*SD*	Skew	Kurtosis	*SE*
Q1 Wonder pass	2.35	0.80	0.13	−0.48	0.09
Q5 Get a bad result	2.06	0.90	0.46	−0.66	0.10
Q6 Hard to remember	2.13	0.84	0.35	−0.52	0.09
Q9 Worry fail	2.16	0.84	0.40	−0.42	0.09
Q11 Worry do wrong	2.59	0.94	0.30	−1.07	0.10
Q13 What result will be	2.45	1.02	0.04	−1.14	0.11
Q15 Answers right	2.76	0.85	−0.22	−0.65	0.09
Q16 Studied more	2.01	0.96	0.70	−0.44	0.10
Q19 Answers wrong	1.93	0.88	0.75	−0.13	0.10
Q21 How hard test	2.34	0.93	0.33	−0.77	0.10
Q24 Happen if fail	1.95	0.97	0.70	−0.58	0.11
Q27 How badly	1.79	0.77	0.83	0.43	0.08
Q29 Parents say	1.81	1.12	0.98	−0.61	0.12

Table G.6

Mean, Skew and Kurtosis for the Original Thoughts' Subscale, Post-test, Intervention Group (n = 85)

Question	*M*	*SD*	Skew	Kurtosis	*SE*
Q1 Wonder pass	2.46	0.85	0.35	−0.60	0.09
Q5 Get a bad result	1.82	0.86	0.78	−0.20	0.09
Q6 Hard to remember	1.86	0.79	0.54	−0.44	0.09
Q9 Worry fail	1.85	0.88	0.81	−0.12	0.10
Q11 Worry do wrong	2.28	0.89	0.41	−0.56	0.10
Q13 What result will be	2.47	1.03	0.14	−1.16	0.11
Q15 Answers right	2.52	0.91	−0.05	−0.83	0.10
Q16 Studied more	1.84	0.94	1.01	0.16	0.10
Q19 Answers wrong	1.68	0.88	1.18	0.57	0.09
Q21 How hard test	2.18	0.97	0.43	−0.80	0.10
Q24 Happen if fail	1.93	1.04	0.76	−0.72	0.11
Q27 How badly	1.60	0.80	1.37	1.45	0.09
Q29 Parents say	1.81	0.99	1.02	−0.10	0.11

Table G.7

Mean, Skew and Kurtosis for the Original Thoughts' Subscale Pre-test, Control Group (n = 85)

Question	*M*	*SD*	Skew	Kurtosis	*SE*
Q1 Wonder pass	2.41	0.93	0.17	−0.86	0.10
Q5 Get a bad result	2.01	0.92	0.89	0.10	0.10
Q6 Hard to remember	2.13	1.02	0.54	−0.85	0.11
Q9 Worry fail	2.22	1.07	0.47	−1.05	0.12
Q11 Worry do wrong	2.41	0.95	0.25	−0.90	0.10
Q13 What result will be	2.55	1.10	0.03	−1.35	0.12
Q15 Answers right	2.88	0.99	−0.27	−1.18	0.11
Q16 Studied more	2.34	1.04	0.42	−1.03	0.11
Q19 Answers wrong	2.18	1.06	0.54	−0.94	0.11
Q21 How hard test	2.33	0.96	0.45	−0.76	0.10
Q24 Happen if fail	2.04	1.13	0.64	−1.03	0.12
Q27 How badly	1.95	1.02	0.89	−0.35	0.11
Q29 Parents say	2.13	1.11	0.57	−1.05	0.12

Table G.8

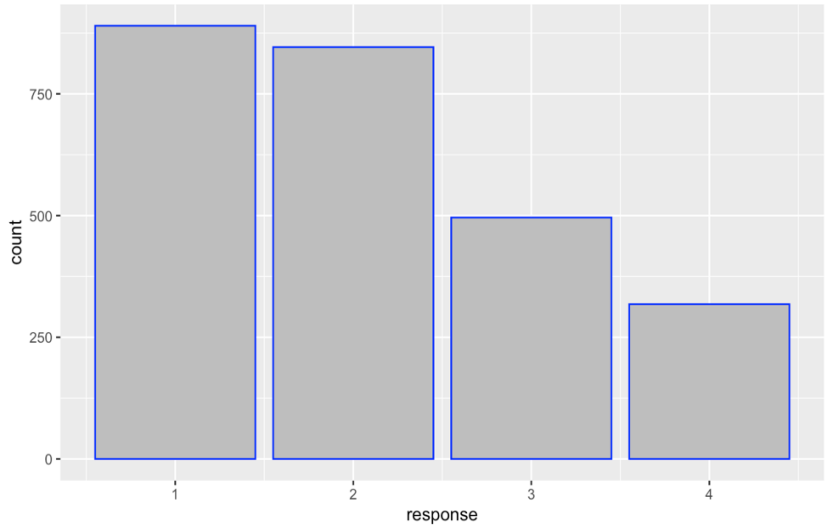
Mean, Skew and Kurtosis for the Original Thoughts' Subscale, Post-test, Control Group (n =85)

Question	*M*	*SD*	Skew	Kurtosis	*SE*
Q1 Wonder pass	2.39	0.96	0.20	−0.95	0.10
Q5 Get a bad result	2.04	0.92	0.84	0.02	0.10
Q6 Hard to remember	2.09	0.88	0.64	−0.18	0.10
Q9 Worry fail	1.91	0.91	0.84	−0.05	0.10
Q11 Worry do wrong	2.41	0.92	0.30	−0.77	0.10
Q13 What result will be	2.46	1.05	0.17	−1.21	0.11
Q15 Answers right	2.55	1.04	−0.05	−1.20	0.11
Q16 Studied more	1.95	0.92	0.72	−0.36	0.10
Q19 Answers wrong	1.92	0.97	0.79	−0.41	0.10
Q21 How hard test	2.31	1.00	0.36	−0.95	0.11
Q24 Happen if fail	1.98	1.09	0.70	−0.92	0.12
Q27 How badly	1.75	0.87	1.02	0.31	0.09
Q29 Parents say	2.06	1.13	0.58	−1.13	0.96

Normal Distribution for the Original Thoughts' Subscale from the Children's Test Anxiety Scale Instrument.

Figure G.1

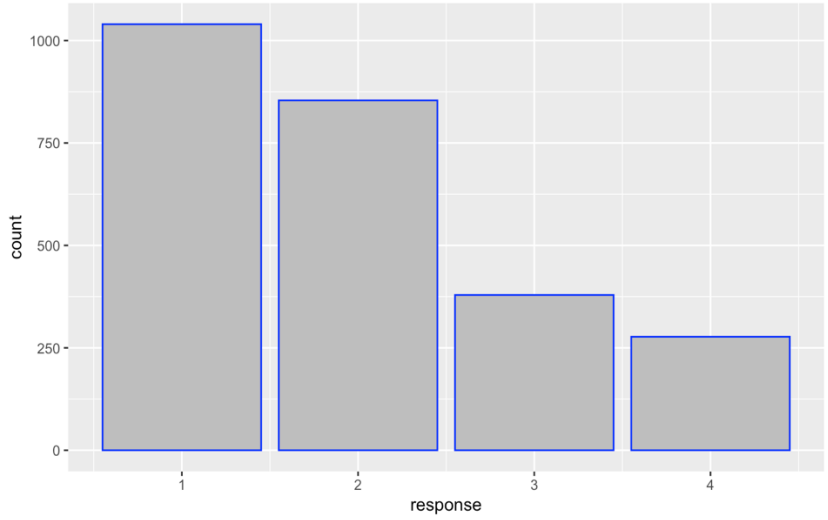
Responses for the Original Thoughts' Subscale Pre-test Intervention Group, n =85



1 = Never
2 = Sometimes
3 = Many Times
4 = Always

Figure G.2

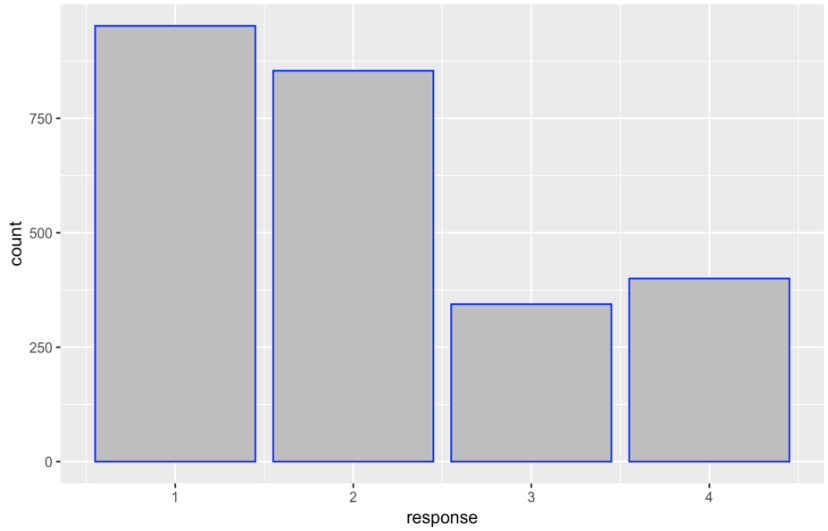
Responses for the Original Thoughts' Subscale Post-test Intervention Group, n = 85



1 = Never
2 = Sometimes
3 = Many Times
4 = Always

Figure G.3

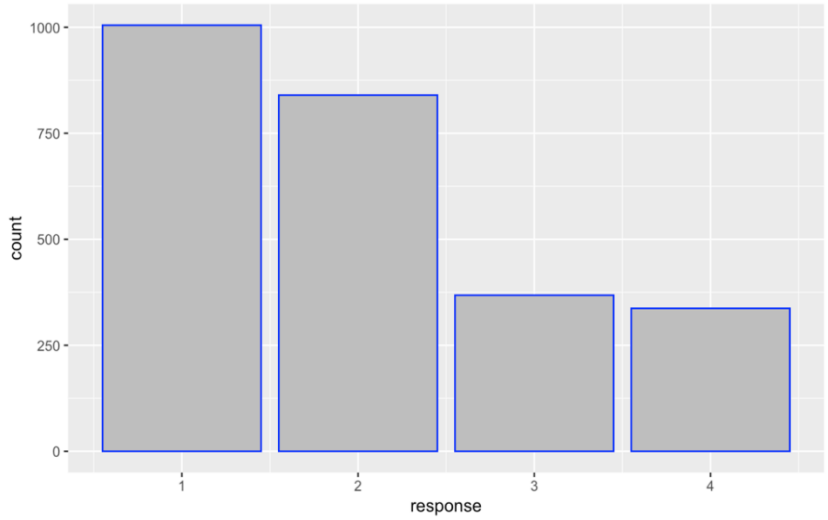
Responses for the Original Thoughts' Subscale Pre-test Control Group, n = 85



1 = Never
2 = Sometimes
3 = Many Times
4 = Always

Figure G.4

Responses for the Original Thoughts' Subscale, Post-test, Control Group, n =85



1 = Never
2 = Sometimes
3 = Many Times
4 = Always

Figure G.5

Violin Plot - Responses for Adapted Thoughts' Subscale for Intervention Pre-test, n = 85

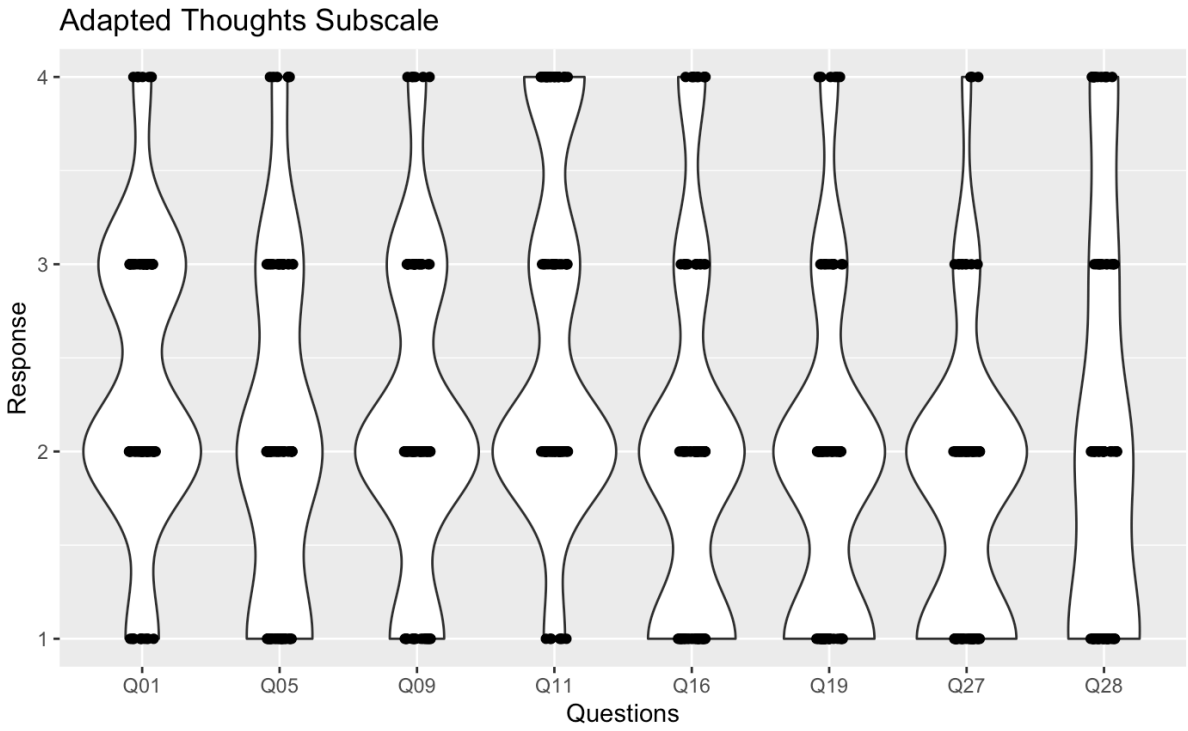


Figure G.6

Violin Plot – Responses for Adapted Thoughts' Subscale for Intervention Post -test, n = 85

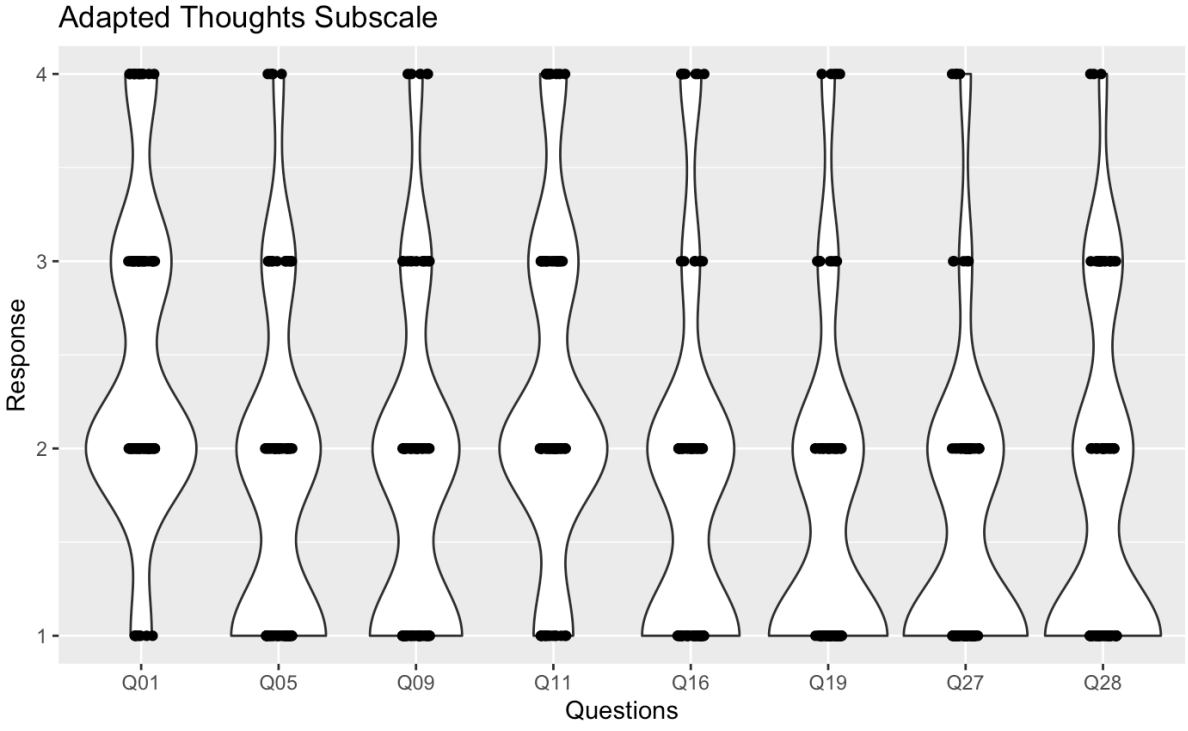


Figure G.7

Violin Plot – Responses for Adapted Thoughts' Subscale for Control Pre-test, n = 85

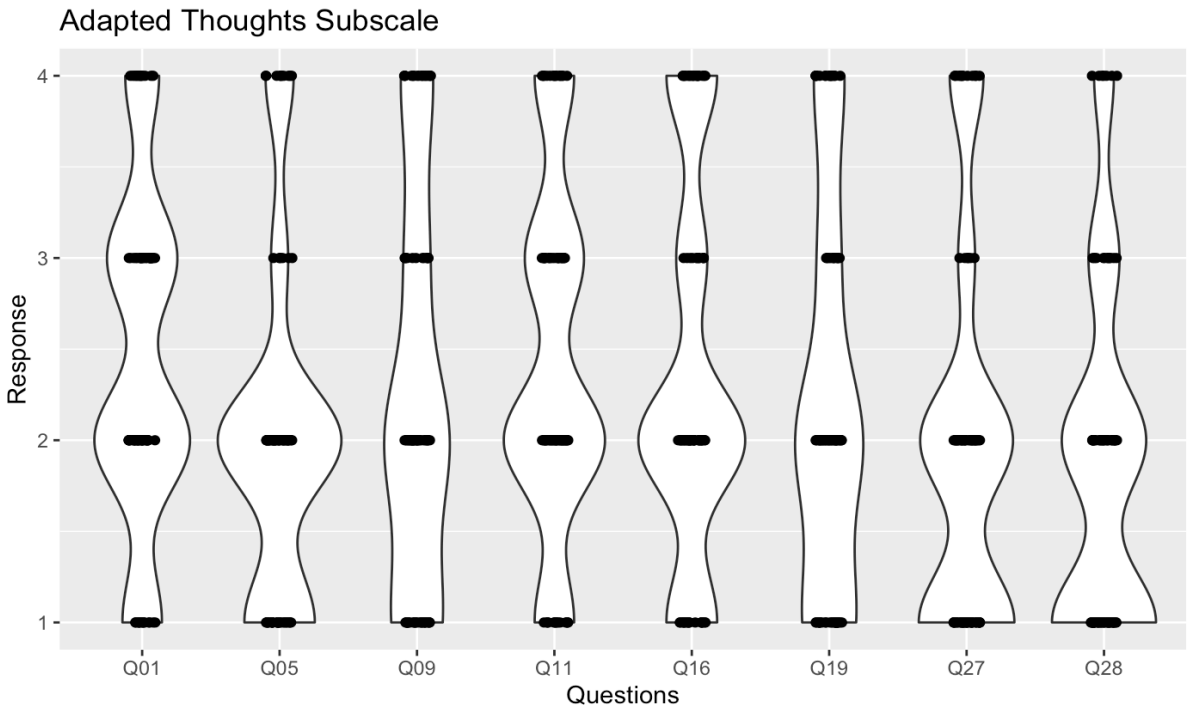


Figure G.8

Violin Plot Responses for Adapted Thoughts' Subscale for Control Post-test, n = 85

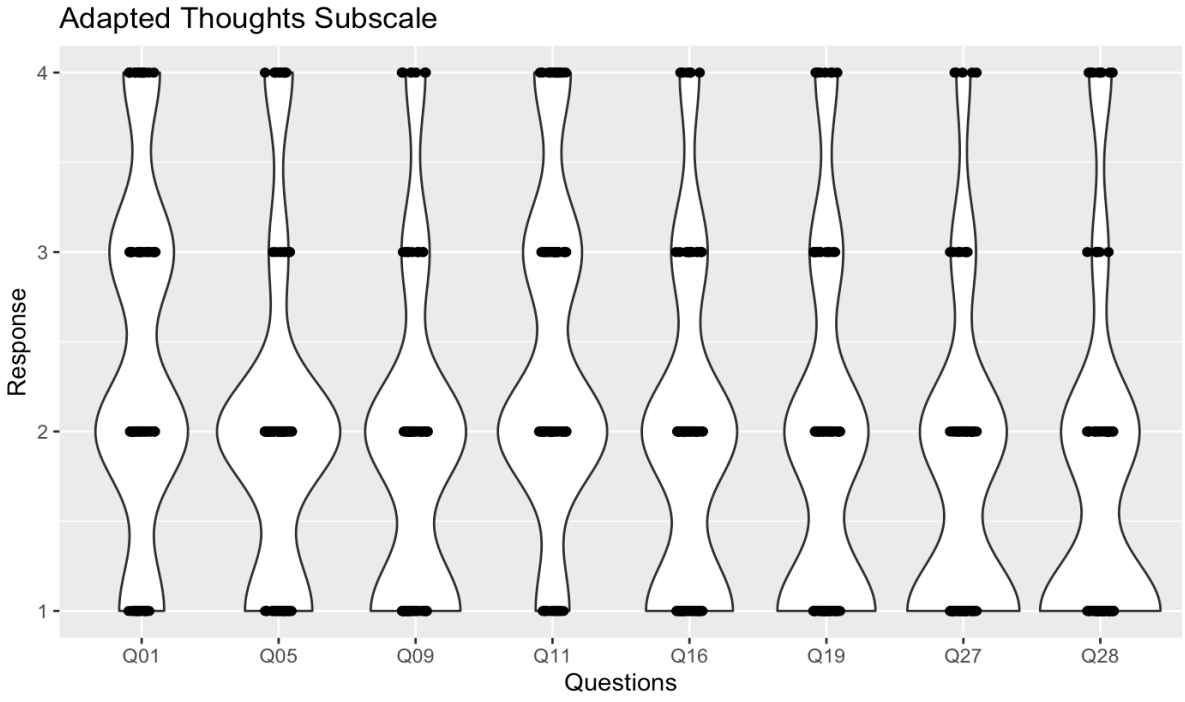
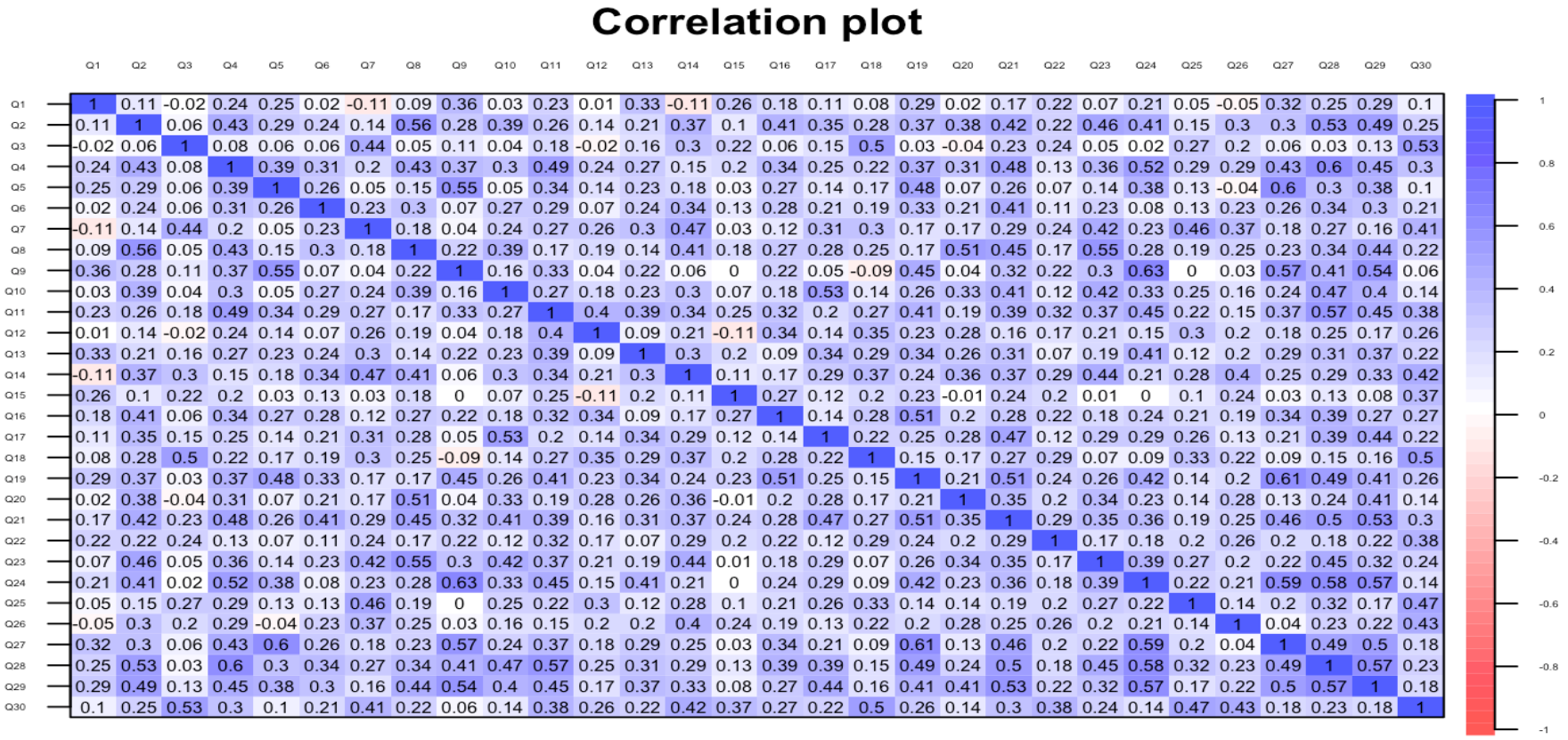


Figure G.9

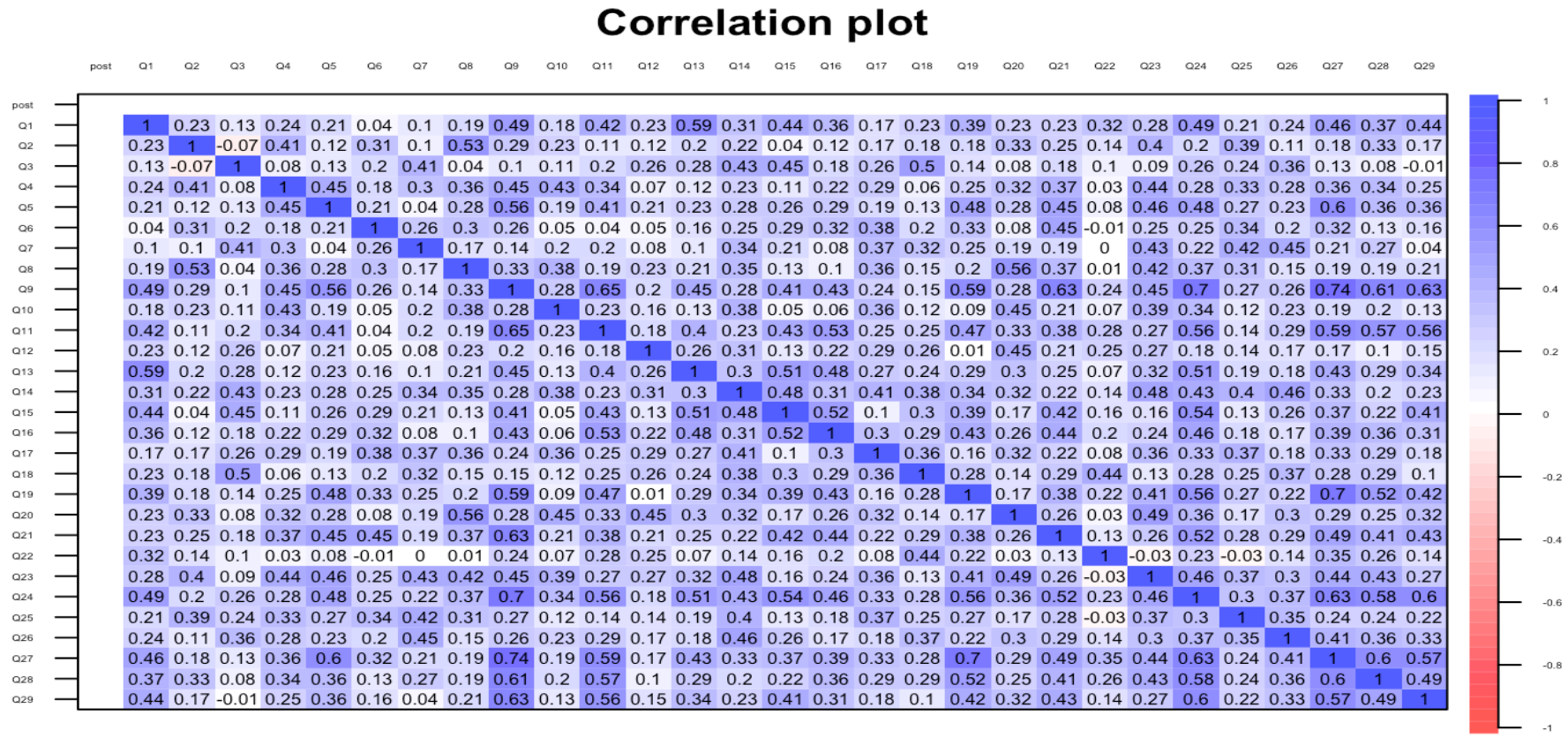
Children's Test Anxiety Scale Pre-test Intervention Group Correlations, n = 85



Q1 will I pass Q2 heart-beat Q3 look room Q4 nervous Q5 bad result	Q6 remember answers Q7 play pen /pencil Q8 face hot Q9 worry fail Q10 belly funny	Q11 worry wrong Q12 check time Q13 result will be Q14 hard sit still Q15 answers right	Q16 studied more Q17 head hurts Q18 look people Q19 answers wrong Q20 feel warm	Q21worry how hard Q22 finish quickly Q23 hand shakes Q24 happen fail Q25 go bathroom	Q26 tap feet Q27 how badly Q28 feel scared Q29 parents say Q30 stare

Figure G.10

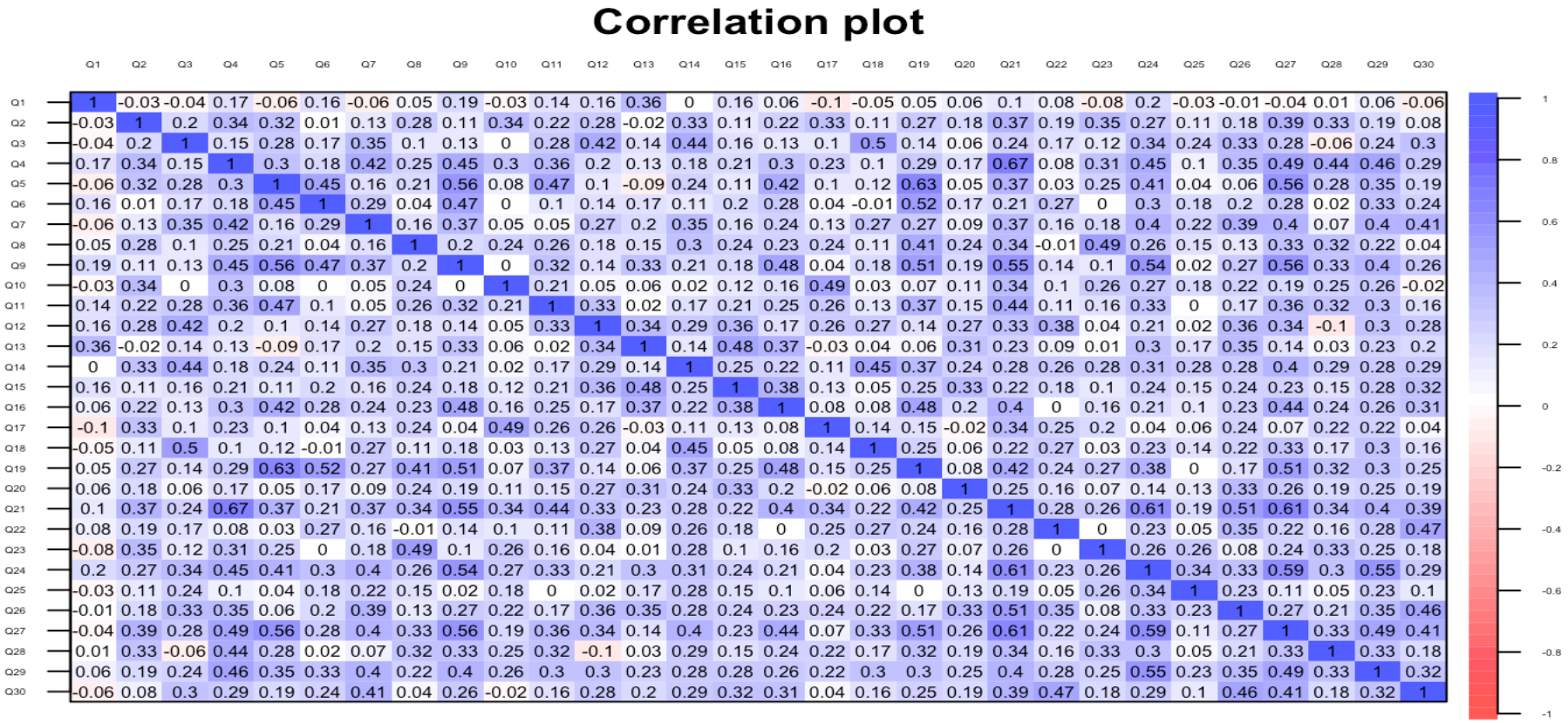
Children's Test Anxiety Scale Post-test Intervention Group Correlations, n = 85



Q1 will I pass	Q6 remember answers	Q11 worry wrong	Q16 studied more	Q21worry how hard	Q26 tap feet
Q2 heart-beat	Q7 play pen /pencil	Q12 check time	Q17 head hurts	Q22 finish quickly	Q27 how badly
Q3 look room	Q8 face hot	Q13 result will be	Q18 look people	Q23 hand shakes	Q28 feel scared
Q4 nervous	Q9 worry fail	Q14 hard sit still	Q19 answers wrong	Q24 happen fail	Q29 parents say
Q5 bad result	Q10 belly funny	Q15 answers right	Q20 feel warm	Q25 go bathroom	Q30 stare

Figure G.11

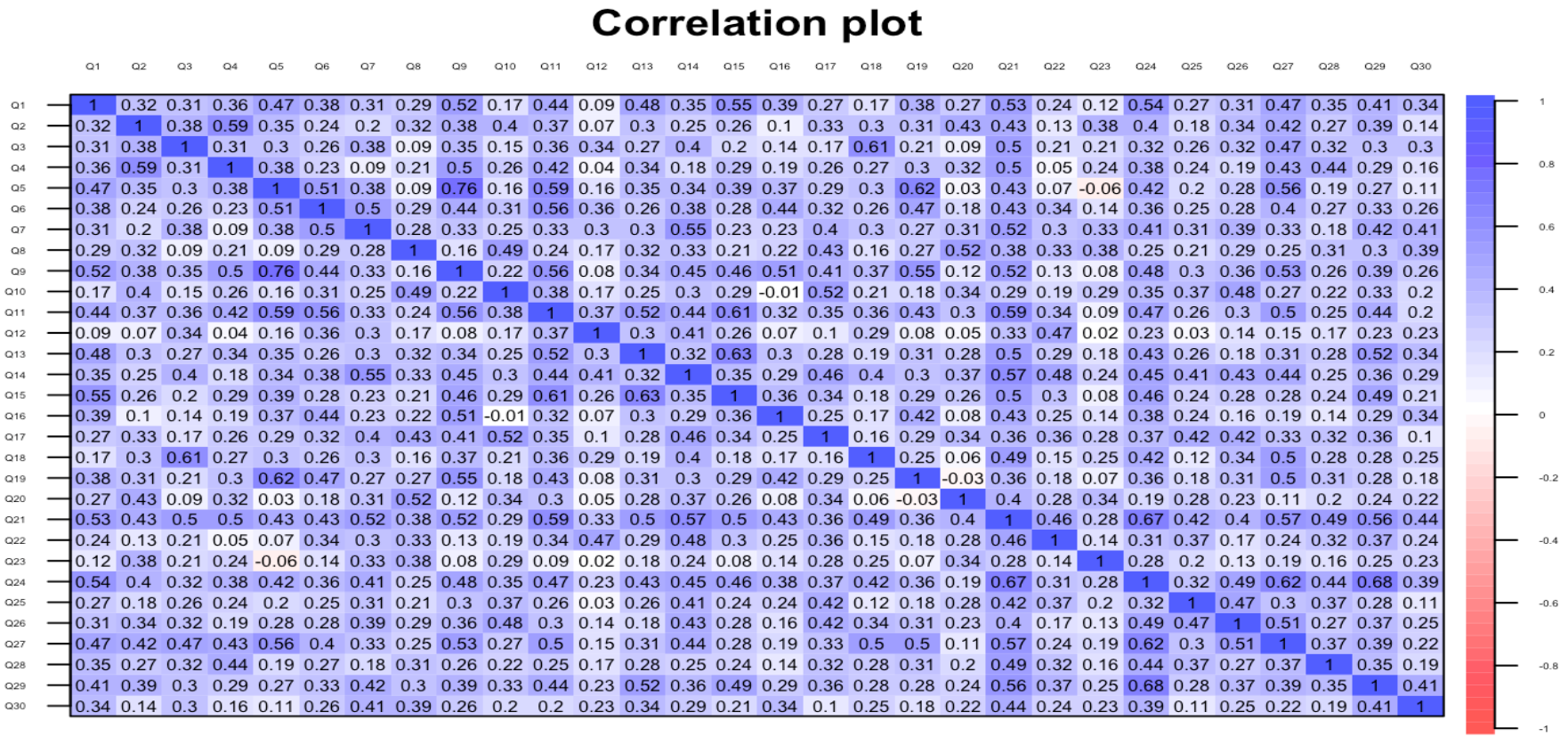
Children's Test Anxiety Pre-test Control Group Correlations, n = 85



Q1 will I pass Q2 heart-beat Q3 look room Q4 nervous Q5 bad result	Q6 remember answers Q7 play pen /pencil Q8 face hot Q9 worry fail Q10 belly funny	Q11 worry wrong Q12 check time Q13 result will be Q14 hard sit still Q15 answers right	Q16 studied more Q17 head hurts Q18 look people Q19 answers wrong Q20 feel warm	Q21worry how hard Q22 finish quickly Q23 hand shakes Q24 happen fail Q25 go bathroom	Q26 tap feet Q27 how badly Q28 feel scared Q29 parents say Q30 stare

Figure G.12

Children's Test Anxiety Post-test Control Group Correlations, n = 85



Q1 will I pass Q2 heart-beat Q3 look room Q4 nervous Q5 bad result	Q6 remember answers Q7 play pen /pencil Q8 face hot Q9 worry fail Q10 belly funny	Q11 worry wrong Q12 check time Q13 result will be Q14 hard sit still Q15 answers right	Q16 studied more Q17 head hurts Q18 look people Q19 answers wrong Q20 feel warm	Q21worry how hard Q22 finish quickly Q23 hand shakes Q24 happen fail Q25 go bathroom	Q26 tap feet Q27 how badly Q28 feel scared Q29 parents say Q30 stare

Initial Confirmatory Factor Analysis

Table G.9

Shapiro-Wilk Test for Normality of Distribution for Pre and Post Timepoints and Intervention and Control Group (N = 340)

Timepoint	Group	*Shapiro-Wilk*	
		Statistic	*p*
Pre-test	intervention	0.84	< .001
	control	0.82	< .001
Post-test	intervention	0.81	< .001
	control	0.82	< .001

Note. Children's Test Anxiety Scale Instrument.

Table G.10

Kaiser-Meyer-Olkin Measures of Sampling Adequacy (N = 340)

Timepoint	Children's Test Anxiety Scale	
	Intervention	Control
Pre-test	0.80	0.78
Post-test	0.76	0.85

Note. Children's Test Anxiety Scale Instrument

Table G.11

Bartlett's Test for Sphericity for the Children's Test Anxiety Scale instrument (N = 340)

Timepoint	Group	X^2	*df*	*p*
Pre-test	Intervention	1164.17	435	< .001
	Control	1128.74	435	< .001
Post-test	Intervention	1363.72	435	< .001
	Control	1353.03	435	< .001

Note. Children's Test Anxiety Scale

Table G.12

Determinant Matrix Values for the Groups and Timepoints (N = 340)

Timepoint	Group	*R*
Pre-test	Intervention	1.23 e−09
	Control	199.0 e−09
Post-test	Intervention	8.04 e−09
	Control	9.31 e−09

Note. Children's Test Anxiety Scale Data

Figure G.13

Children's Test Anxiety Scale Intervention Pre-test Correlations, (X = Significant Values)

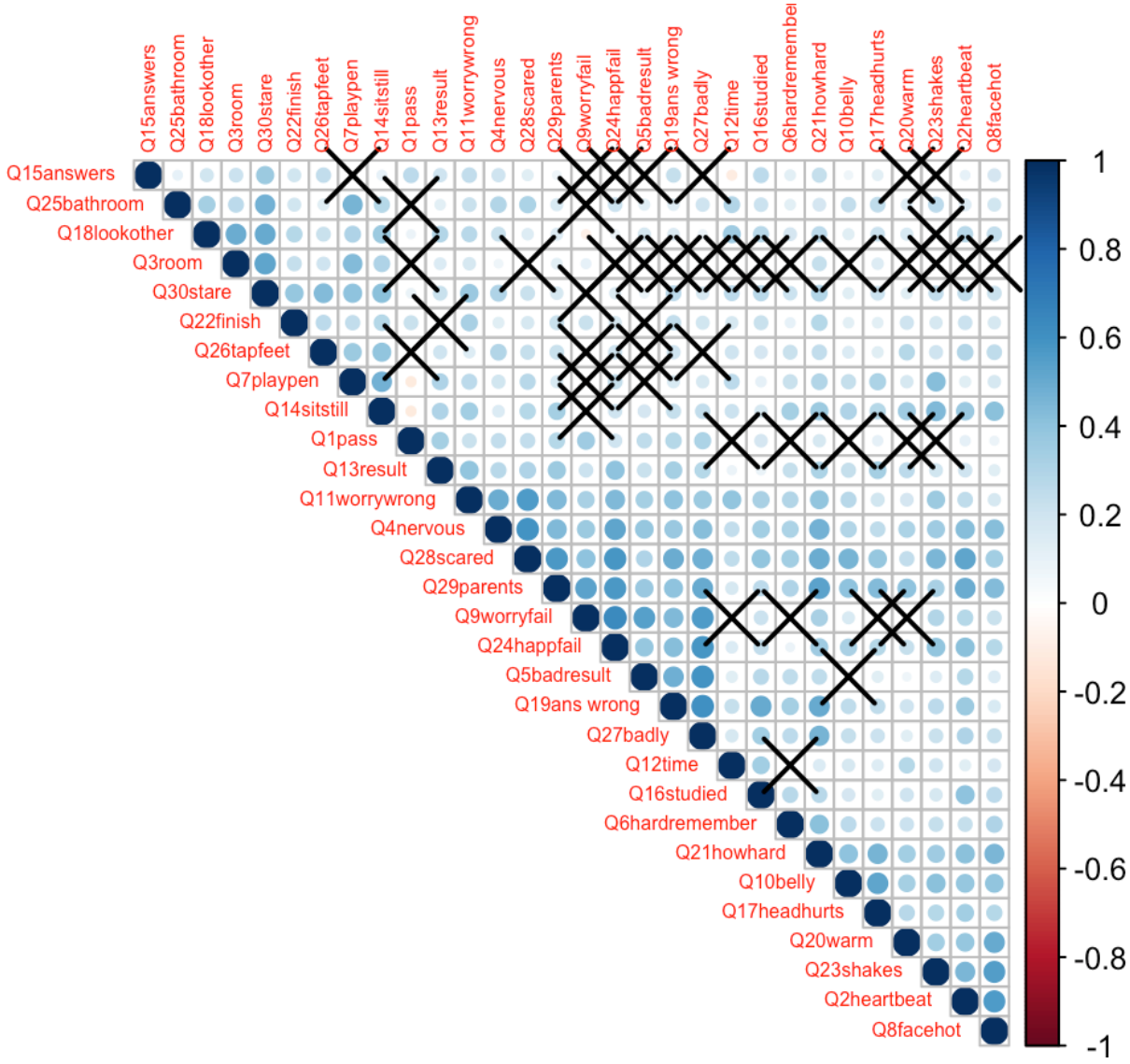


Figure G.14

Children's Test Anxiety Scale Intervention Post-test Correlations (X = Significant Values)

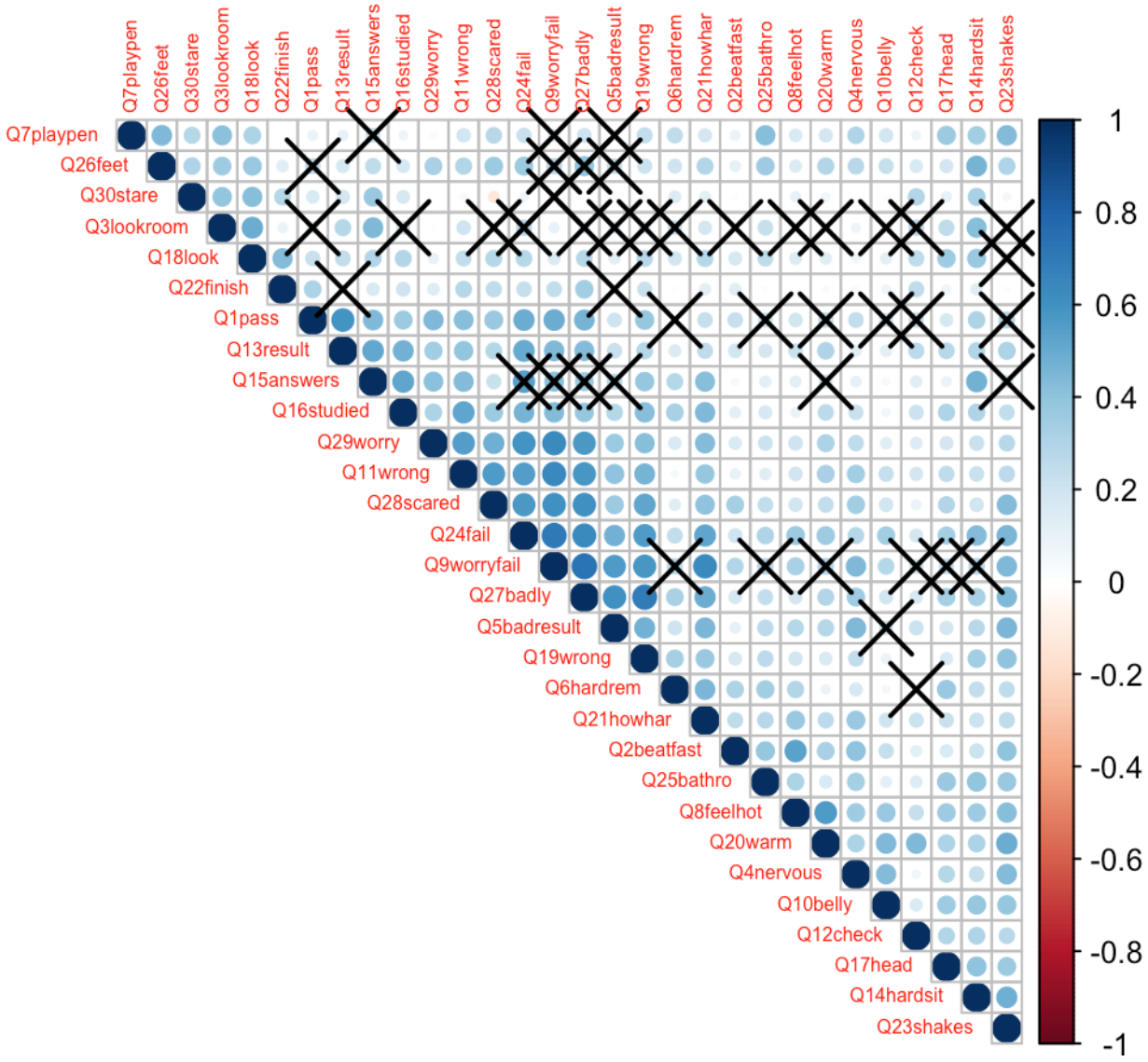


Figure G.15

Children's Test Anxiety Scale Control Pre-test Correlations (X = Significant Values)

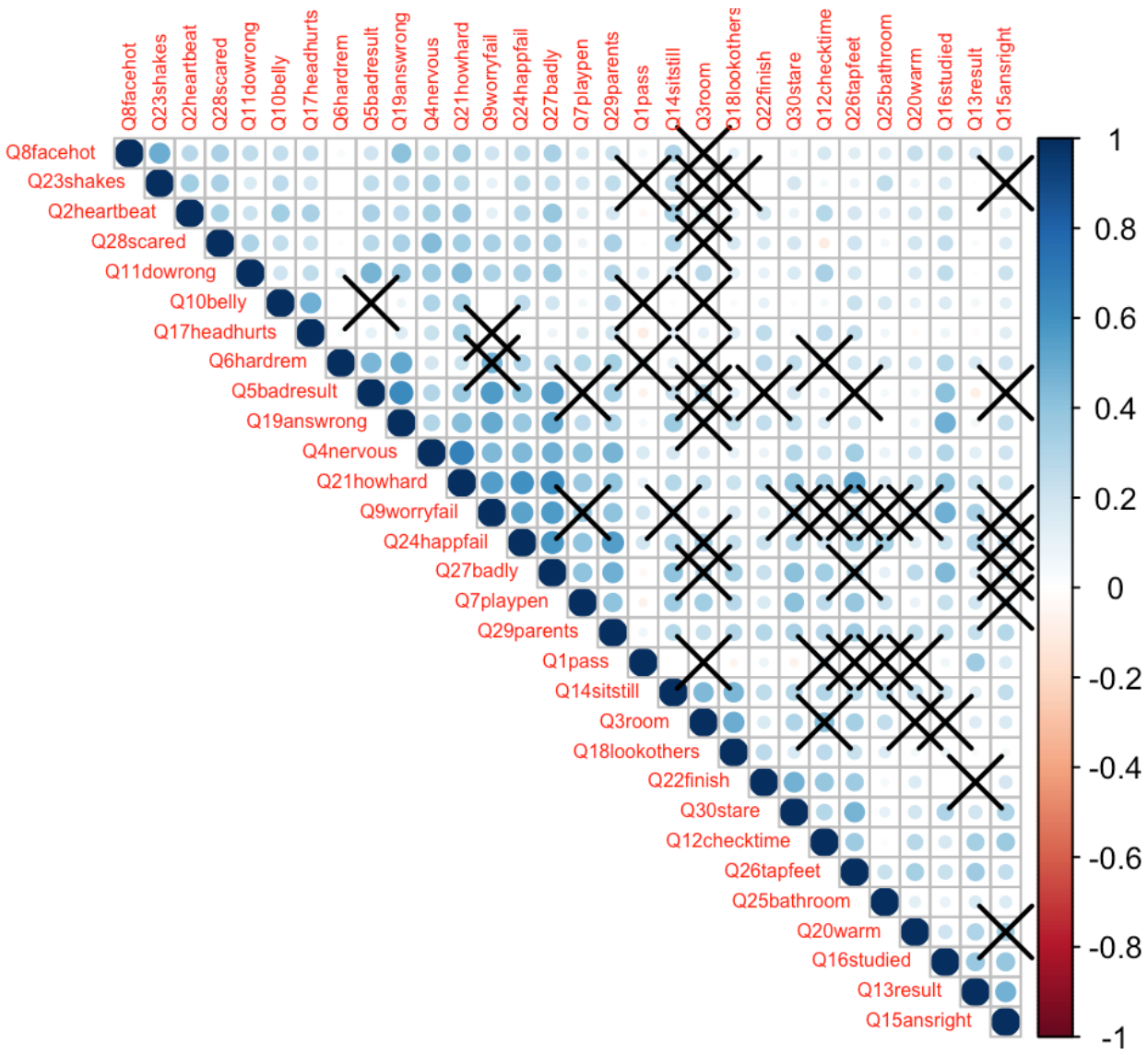


Figure G.16

Children's Test Anxiety Scale Control Post-test (X = Significant Values)

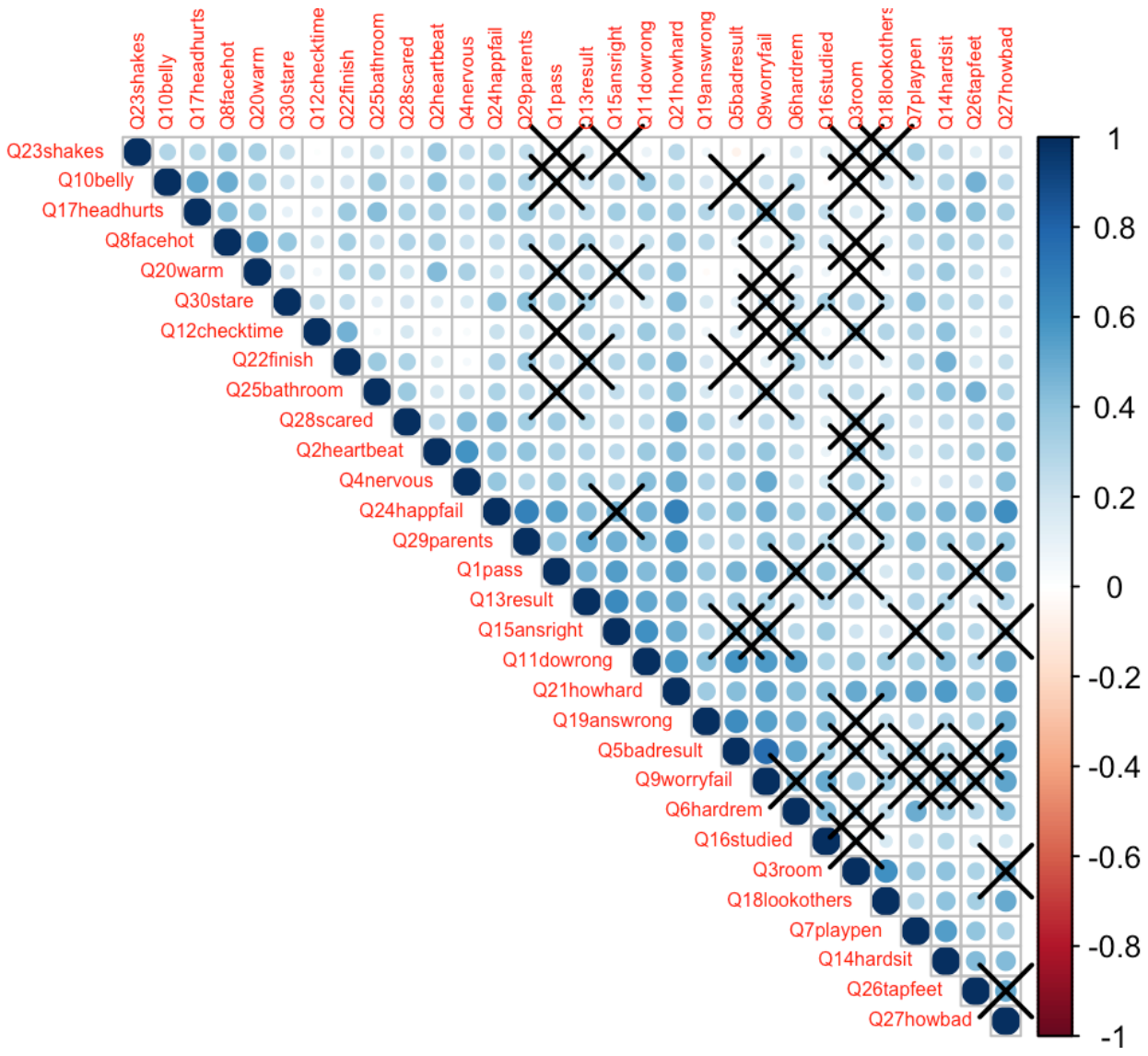


Table G.13

Factor Loadings from the Original Children's Test Anxiety Scale Development Paper, Wren and Benson, (2004)

Question	Subscale		
	Autonomic Reactions	Thoughts	Off Task Behaviours
Q1 Wonder if pass		0.65	
Q2 Heart beat fast	0.69		
Q3 Look around room			0.74
Q4 Feel nervous	0.69		
Q5 Get a bad result		0.76	
Q6 Hard to remember		0.37	
Q7 Play pen /pencil			0.42
Q8 Face feels hot	0.62		
Q9 Worry fail		0.70	
Q10 Belly funny	0.57		
Q11 Worry do wrong		0.68	
Q12 Check the time			0.55
Q13 What result will be		0.63	
Q14 Hard sit still			0.54
Q15 Answers right		0.49	
Q16 Studied more		0.48	
Q17 Head hurts	0.63		
Q18 Look at others			0.71
Q19 Answers wrong		0.60	
Q20 Feel warm	0.61		
Q21 How hard test		0.55	
Q22 Finish quickly			0.40
Q23 Hand shakes	0.60		
Q24 Happen if fail		0.75	
Q25 Go to bathroom	0.41		
Q26 Tap my feet			0.53
Q27 How badly		0.58	
Q28 Feel scared	0.67		
Q29 Parents say		0.72	
Q30 I stare			0.60

Exploratory Factor Analysis on the Children's Test Anxiety Scale

Table G.14

***Condition 1**: Exploratory Factor Analysis Showing the Factor Loadings for **Three** Factors with Pre-test Intervention Group (n = 85)*

Question	Thoughts	Autonomic Reactions	Off Task Behaviours
Q1 Wonder if pass	**0.53 (**0.65)	−0.19	
Q2 Heart beat fast	0.13	0.65 (0.69)	
Q3 Look around room		−0.31	**0.80** (0.74)
Q4 Feel nervous	**0.40**	0.33 (0.69)	
Q5 Get a bad result	**0.74** (0.76)	−0.17	
Q6 Hard to remember	(0.37)	0.32	**0.52**
Q7 Play pen /pencil	−0.14	0.20	**0.52 (**0.42)
Q8 Face feels hot		**0.80** (0.62)	−0.10
Q9 Worry fail	**0.86** (0.70)		−0.19
Q10 Belly funny		**0.72** (0.57)	−0.14
Q11 Worry do wrong	**0.41** (0.68)		0.30
Q12 Check the time		0.18	0.24 (0.55)
Q13 What result will be	0.27 (0.63)	0.13	0.19
Q14 Hard sit still	−0.16	**0.44**	0.38 (0.54)
Q15 Answers right	(0.49)	−0.13	**0.43**
Q16 Studied more	0.30 (0.48)		0.20
Q17 Head hurts		**0.53** (0.63)	
Q18 Look at others	−0.12		**0.71** (0.71)
Q19 Answers wrong	**0.63** (0.60)		0.11
Q20 Feel warm	−0.16	**0.76** (0.61)	−0.13
Q21 How hard test	0.26 (0.55)	**0.43**	0.12
Q22 Finish quickly	0.12		**0.42** (0.40)
Q23 Hand shakes		**0.69** (0.60)	
Q24 Happen if fail	**0.64** (0.75)	0.27	−0.15
Q25 Go to bathroom		(0.41)	**0.49**
Q26 Tap my feet	−0.17	0.28	0.38 (0.53)
Q27 How badly	**0.80** (0.58)		
Q28 Feel scared	**0.44**	**0.47** (0.67)	
Q29 Parents say	**0.47**	**0.47** (0.72)	−0.13
Q30 I stare		−0.14	**0.90** (0.60)

Note. Brackets () denote the original loadings. **Bold =** the values considered as loadings here.

Table G.15

***Condition 2**: Exploratory Factor Analysis Showing the Factor Loadings for **Four** Factors With the Pre-test Intervention Group (n = 85)*

Question	Thoughts	Autonomic Reactions	Off Task Behaviours	Factor 4
Q1 Wonder if pass	**0.42** (0.65)	−0.117	−0.19	0.29
Q2 Heart beat fast		**0.73** (0.69)	−0.19	0.16
Q3 Look around room		−0.36 (0.74)	**0.74**	0.16
Q4 Feel nervous	0.36	0.34 (0.69)		
Q5 Get a bad result	**0.71** (0.76)	−0.16		
Q6 Hard to remember	(0.37)	0.34		
Q7 Play pen /pencil			**0.92** (0.42)	−0.38
Q8 Face feels hot	−0.17	**0.89** (0.62)	−0.19	0.12
Q9 Worry fail	**0.92** (0.70)	−0.12	−0.11	−0.13
Q10 Belly funny		**0.64** (0.57)		−0.13
Q11 Worry do wrong	**0.42** (0.68)		0.25	0.12
Q12 Check the time		0.16	0.23 (0.55)	
Q13 What result will be	0.31 (0.63)		0.24	
Q14 Hard sit still	−0.11	0.38	**0.45** (0.54)	
Q15 Answers right	−0.13 (0.49)			**0.56**
Q16 Studied more	0.16 (0.48)	0.19		0.38
Q17 Head hurts		**0.45** (0.63)	0.14	
Q18 Look at others	−0.21		**0.47** (0.71)	**0.41**
Q19 Answers wrong	**0.57** (0.60)			0.20
Q20 Feel warm	−0.20	**0.78** (0.61)	−0.13	
Q21 How hard test	0.23 (0.55)	**0.41**		0.14
Q22 Finish quickly			0.29 (0.40)	0.23
Q23 Hand shakes	0.11	**0.59** (0.60)	0.16	−0.28
Q24 Happen if fail	**0.78** (0.75)	0.14		−0.27
Q25 Go to bathroom		(0.41)		
Q26 Tap my feet			**0.56** (0.53)	
Q27 How badly	**0.83** (0.53)			
Q28 Feel scared	**0.48**	0.39 (0.67)		
Q29 Parents say	**0.48** (0.72)	**0.44**	−0.11	
Q30 I stare			**0.66** (0.60)	**0.40**

Note. Brackets () denote the original loadings. **Bold =** the values considered as loadings here.

Table G.16

***Condition 3**: Exploratory Factor Analysis Showing the Factor Loadings for **Three** Factors with Post-test Intervention Group (n = 85)*

Question	Thoughts	Autonomic Reactions	Off Task Behaviours
Q1 Wonder if pass	**0.54** (0.65)	0.13	0.23
Q2 Heart beat fast	0.12	**0.55**	
Q3 Look around room			**0.69**
Q4 Feel nervous	0.29	**0.56** (0.69)	
Q5 Get a bad result	**0.53** (0.76)	0.35	
Q6 Hard to remember	0.19 (0.37)	0.30	0.23
Q7 Play pen /pencil		0.39	**0.41**
Q8 Face feels hot	0.12	**0.65**	
Q9 Worry fail	**0.84** (0.70)	0.32	−0.24
Q10 Belly funny	0.11	**0.54**	
Q11 Worry do wrong	**0.71** (0.68)	0.15	0.11
Q12 Check the time	0.11	0.26	0.31
Q13 What result will be	**0.49** (0.63)	0.13	0.33
Q14 Hard sit still	0.20	**0.44**	**0.56**
Q15 Answers right	**0.53** (0.49)		**0.55**
Q16 Studied more	**0.53** (0.48)		0.31
Q17 Head hurts	0.14	**0.49**	0.29
Q18 Look at others	0.19	0.13	**0.60**
Q19 Answers wrong	**0.65** (0.60)	0.20	0.13
Q20 Feel warm	0.19	**0.57**	0.12
Q21 How hard test	**0.56** (0.55)	0.29	0.15
Q22 Finish quickly	0.31		0.23
Q23 Hand shakes	0.29	**0.69**	
Q24 Happen if fail	**0.72** (0.75)	0.31	0.23
Q25 Go to bathroom	0.14	**0.49**	0.21
Q26 Tap my feet	0.25	0.31	**0.40**
Q27 How badly	**0.78** (0.58)	0.27	
Q28 Feel scared	**0.63**	(0.67) 0.31	
Q29 Parents say	**0.69** (0.72)	0.16	
Q30 I stare			**0.68**

Note. Brackets () denote the original loadings. **Bold =** the values considered as loadings here.

Table G.17

***Condition 4**: Exploratory Factor Analysis Showing the Factor Loadings for **Four** Factors Post-test Intervention Group (n = 85)*

Question	Thoughts	Autonomic Reactions	Off Task Behaviours	Factor 4
Q1 Wonder if pass	**0.45** (0.65)	0.11	0.29	−0.16
Q2 Heart beat fast		**0.59**	−0.15	
Q3 Look around room	−0.13	−0.17	**0.60**	**0.46**
Q4 Feel nervous	0.26	0.37	−0.23	0.19
Q5 Get a bad result	**0.58** (0.76)	**0.113**	−0.12	
Q6 Hard to remember	0.13			0.30
Q7 Play pen /pencil			0.12	**0.74**
Q8 Face feels hot	−0.11	**0.84**		
Q9 Worry fail	**0.92** (0.70)		−0.10	−0.10
Q10 Belly funny		**0.56**		
Q11 Worry do wrong	**0.73** (0.68)			
Q12 Check the time	−0.12	**0.41**	0.32	
Q13 What result will be	0.33	0.18	**0.43**	−0.19
Q14 Hard sit still		0.28	**0.41**	0.34
Q15 Answers right	0.34	−0.11	**0.66**	
Q16 Studied more	**0.44** (0.48)		0.35	
Q17 Head hurts		0.32		0.33
Q18 Look at others		−0.12	**0.45**	**0.44**
Q19 Answers wrong	**0.75** (0.60)	−0.21		0.22
Q20 Feel warm		**0.77**		
Q21 How hard test	**0.51** (0.55)	0.13		
Q22 Finish quickly	0.30	−0.19	0.22	
Q23 Hand shakes	0.19	**0.49**	−0.13	0.26
Q24 Happen if fail	**0.66** (0.75)	0.13	0.18	
Q25 Go to bathroom		0.20		**0.44**
Q26 Tap my feet	0.15		0.19	**0.48**
Q27 How badly	**0.89**(0.58)	−0.14		0.15
Q28 Feel scared	**0.74**(0.67)		−0.18	0.18
Q29 Parents say	**0.73** (0.72)			−0.19
Q30 I stare	−0.36		**0.68**	0.25

Note. Brackets () denote the original loadings. **Bold =** the values considered as loadings here.

Table G.18

Condition 5*: Exploratory Factor Analysis Showing the Factor Loadings for* ***Three*** *Factors*

Pre-test Control Group (n = 85)

Question	Thoughts	Autonomic Reactions	Off Task Behaviours
Q1 Wonder if pass	0.29 (0.65)		−0.20
Q2 Heart beat fast		**0.61** (0.69)	
Q3 Look around room			0.35 (0.74)
Q4 Feel nervous	0.13	**0.40** (0.69)	0.28
Q5 Get a bad result	**1.02** (0.76)		−0.38
Q6 Hard to remember	**0.61** (0.37)	−0.33	0.17
Q7 Play pen /pencil			**0.52** (0.42)
Q8 Face feels hot		**0.44** (0.62)	
Q9 Worry fail	**0.70** (0.70)	−0.16	0.22
Q10 Belly funny	−0.22	**0.65** (0.57)	
Q11 Worry do wrong	0.36 (0.68)	0.31	
Q12 Check the time	−0.15		**0.58** (0.55)
Q13 What result will be	−0.18 (0.63)	−0.26	**0.77**
Q14 Hard sit still		0.20	0.32 (0.54)
Q15 Answers right	(0.49)		**0.52**
Q16 Studied more	**0.44** (0.48)		0.20
Q17 Head hurts	−0.20	**0.60** (0.63)	
Q18 Look at others		0.12	0.29 (0.71)
Q19 Answers wrong	**0.76** (0.60)		
Q20 Feel warm	−0.10	(0.61)	**0.45**
Q21 How hard test	0.17 (0.55)	0.39	**0.41**
Q22 Finish quickly	−0.12		**0.47** (0.40)
Q23 Hand shakes		**0.53** (0.57)	
Q24 Happen if fail	0.30 (0.75)	0.16	0.38
Q25 Go to bathroom	−0.14	0.14 (0.41)	0.30
Q26 Tap my feet	−0.25	0.15	**0.74** (0.53)
Q27 How badly	**0.50** (0.58)	0.21	0.21
Q28 Feel scared	0.17	**0.46** (0.67)	
Q29 Parents say	0.18 (0.72)	0.18	**0.40**
Q30 I stare			**0.56** (0.60)

Note. Brackets () denote the original loadings. **Bold =** the values considered as loadings here.

Table G.19

***Condition 6**: Exploratory Factor Analysis Showing the Factor Loadings for **Four** Factors*

Pre-test Control Group (n=85)

Question	Thoughts	Autonomic Reactions	Off Task Behaviours	Factor 4
Q1 Wonder if pass	(0.65)		−0.21	**0.47**
Q2 Heart beat fast		**0.58** (0.69)	0.15	−0.26
Q3 Look around room	0.19	−0.24	**0.90** (0.74)	−0.18
Q4 Feel nervous		**0.60** (0.69)	−0.18	0.30
Q5 Get a bad result	**1.03** (0.76)			−0.40
Q6 Hard to remember	**0.57** (0.37)	−0.30		0.20
Q7 Play pen /pencil			0.35 (0.42)	0.30
Q8 Face feels hot		**0.49** (0.62)		
Q9 Worry fail	**0.65** (0.70)		−0.18	0.43
Q10 Belly funny	−0.22	**0.77** (0.57)	−0.17	
Q11 Worry do wrong	0.35 (0.68)	0.31		−0.11
Q12 Check the time	−0.13		**0.55** (0.55)	0.20−
Q13 What result will be	−0.20 (0.63)	−0.12		**0.79**
Q14 Hard sit still			**0.60** (0.54)	
Q15 Answers right	(0.49)		0.14	0.39
Q16 Studied more	**0.40** (0.48)			0.27
Q17 Head hurts	−0.18	**0.60** (0.63)		−0.14
Q18 Look at others			**0.64** (0.71)	
Q19 Answers wrong	**0.69** (0.60)			
Q20 Feel warm	−0.11	0.13 (0.61)		0.35
Q21 How hard test	0.12	**0.56** (0.55)		0.34
Q22 Finish quickly	−0.12		0.37 (0.40)	0.18
Q23 Hand shakes		**0.53** (0.60)		−0.20
Q24 Happen if fail	0.27 (0.75)	0.23	0.12	0.28
Q25 Go to bathroom	−0.12	0.11 (0.41)	0.27	
Q26 Tap my feet	−0.23	0.20	0.31 (0.53)	**0.44**
Q27 How badly	**0.46** (0.58)	0.24	0.18	
Q28 Feel scared	0.13	**0.61** (0.67)	−0.24	
Q29 Parents say	0.17 (0.72)	0.23	0.17	0.23
Q30 I stare			0.37 (0.60)	0.28

Note. Brackets () denote the original loadings. **Bold =** the values considered as loadings here.

Table G.20

***Condition** 7: Exploratory Factor Analysis Showing the Factor Loadings for **Three** Factors*

Post-test Control Group (n = 85)

Question	Thoughts	Autonomic Reactions	Off Task Behaviours
Q1 Wonder if pass	0.39 (0.65)	0.20	0.16
Q2 Heart beat fast	0.19	**0.88** (0.69)	−0.40
Q3 Look around room	0.17	0.22	0.21(0.74)
Q4 Feel nervous	0.36	**0.84** (0.69)	−0.54
Q5 Get a bad result	**1.05 (**0.76)	−0.15	−0.12
Q6 Hard to remember	**0.40** (0.37)	−0.16	**0.45**
Q7 Play pen /pencil			**0.66** (0.42)
Q8 Face feels hot	−0.25	**0.46** (0.62)	0.32
Q9 Worry fail	**0.85** (0.70)		
Q10 Belly funny	−0.11	**0.50** (0.57)	0.14
Q11 Worry do wrong	**0.49** (0.68)		0.26
Q12 Check the time		−0.29	**0.74** (0.55)
Q13 What result will be	0.16 (0.63)	0.20	0.31
Q14 Hard sit still			**0.69** (0.54)
Q15 Answers right	0.28 (0.49)	0.10	0.31
Q16 Studied more	0.38 (0.48)	−0.16	0.33
Q17 Head hurts		0.31 (0.63)	0.26
Q18 Look at others	0.19	0.19	0.19 (0.71)
Q19 Answers wrong	**0.69** (0.60)		
Q20 Feel warm	−0.32	**0.64** (0.61)	0.15
Q21 How hard test	0.16 (0.55)	0.35	**0.44**
Q22 Finish quickly	−0.26		**0.85** (0.40)
Q23 Hand shakes	−0.34	**0.62** (0.60)	
Q24 Happen if fail	0.22 (0.75)	0.32	0.31
Q25 Go to bathroom		0.24 (0.41)	0.32
Q26 Tap my feet	0.10	0.27	0.25 (0.53)
Q27 How badly	**0.49** (0.58)	0.28	
Q28 Feel scared		**0.44** (0.67)	0.12
Q29 Parents say	0.72	0.30	**0.41**
Q30 I stare	−0.11	0.11	**0.49** (0.60)

Note. Brackets () denote the original loadings. **Bold =** the values considered as loadings here.

Table G.21

***Condition 8**: Exploratory Factor Analysis Showing the Factor Loadings for **Four** Factors*

Post-test Control (n = 85)

Question	Thoughts	Autonomic Reactions	Off Task Behaviours	Factor 4
Q1 Wonder if pass	**0.47** (0.65)	0.22		
Q2 Heart beat fast	0.15	**0.73** (0.69)	−0.40	0.20
Q3 Look around room		−0.11	0.23 (0.74)	**0.72**
Q4 Feel nervous	0.35	**0.60** (0.69)	−0.48	0.16
Q5 Get a bad result	**1.12** (0.76)	−0.28	−0.16	
Q6 Hard to remember	**0.46** (0.37)		0.35	
Q7 Play pen /pencil			**0.53** (0.42)	0.14
Q8 Face feels hot	−0.20	**0.69** (0.62)	0.21	−0.16
Q9 Worry fail	**0.91** (0.70)		−0.14	
Q10 Belly funny		**0.62** (0.57)		
Q11 Worry do wrong	**0.55** (0.68)		0.20	
Q12 Check the time	−0.13	−0.27	**0.71** (0.55)	0.22
Q13 What result will be	0.25 (0.63)	0.28	0.26	−0.12
Q14 Hard sit still			**0.57** (0.54)	0.19
Q15 Answers right	**0.41** (0.49)	0.21	0.25	−0.21
Q16 Studied more	**0.49** (0.48)		0.26	−0.15
Q17 Head hurts	0.14	**0.49** (0.63)	0.14	−0.17
Q18 Look at others		−0.15	0.21(0.71)	**0.77**
Q19 Answers wrong	**0.76** (0.60)			
Q20 Feel warm	−0.28	**0.87** (0.61)		−0.20
Q21 How hard test	0.13 (0.55)	0.27	0.37	0.27
Q22 Finish quickly	−0.22	0.12	**0.75** (0.40)	
Q23 Hand shakes	−0.4	**0.63** (0.60)		0.17
Q24 Happen if fail	0.22 (0.75)	0.24	0.24	0.22
Q25 Go to bathroom		0.33 (0.41)	0.23	
Q26 Tap my feet		0.26	0.15 (0.53)	0.21
Q27 How badly	**0.43** (0.58)			**0.43**
Q28 Feel scared		0.35 (0.67)	0.10	0.19
Q29 Parents say	0.34 (0.72)	0.33		
Q30 I stare	−0.12	0.15	**0.43** (0.60)	0.13

Note. Brackets() denote the original loadings. **Bold =** the values considered as loadings here.

Table G.22

Three and Four Factor Loadings and Cronbach's Alpha for the Original Thoughts' Subscale Across All Conditions (N = 340)

Group	Timepoint	Three Factor	Four Factor	Cronbach's *alpha*
Children's Test Anxiety Scale Intervention	Pre-test	Q1, Q4, Q5, Q9, Q11, Q19, Q24, Q27, Q28, Q29	Q1, Q5, Q9, Q11, Q19, Q24, Q27, Q28, Q29	Three Factor = 0.88 Four Factor = 0.87
	Post-test	Q1, Q5, Q9, Q11, Q13, Q15, Q16, Q19, Q21, Q24, Q27, Q28, Q29	Q1, Q5, Q9, Q11, Q16, Q19, Q21, Q24, Q27, Q28, Q29	Three Factor = 0.92 Four Factor = 0.91
Children's Test Anxiety Scale Control	Pre-test	Q5, Q6, Q9, Q16, Q19, Q27	Q5, Q6, Q9, Q16, Q19, Q27	Three Factor = 0.84 Four Factor = 0.84
	Post-test	Q5, Q6, Q9, Q11, Q19, Q27	Q1, Q5, Q6, Q9, Q11, Q15, Q16, Q19, Q27	Three Factor = 0.87 Four Factor = 0.88
Original Thoughts' Subscale		Q1, Q5, Q6, Q9, Q11, Q13, Q15, Q16, Q19, Q21, Q24, Q27, Q29		Alpha = 0.89

Measurement Invariance for the Children's Test Anxiety Adapted Thoughts' Subscale – Timepoint.

Table G.23

Measurement Invariance of Adapted Thoughts' Subscale over Timepoints Pre-test and Post-test (N = 340)

Model	Equal	X^2	*df*	*p*	*RMSEA*	*CFI*	*ΔCFI*	*TLI*	*SRMR*
Configural	Patterns	62.76	56	.00	0.07	0.97		0.95	0.04
Constrained	All	111.74	63	.00	0.07	0.93		0.94	0.08
Metric	Loadings	104.58	56	.00	0.08	0.93	0.00	0.93	0.08
Scalar	Intercepts	100.57	56	.00	0.08	0.94	0.01	0.94	0.08
Strict	Factor	111.54	62	.00	0.07	0.93	0.01	0.94	0.08
Strict	Error	89.72	55	.00	0.07	0.95	0.02	0.95	0.08

Note. ChiSq = chi-square value, df = Degrees of Freedom; RMSEA = root mean square error of approximation; CFI comparative fit indices; TLI = Tucker-Lewis Index; SRMR = Standardised Root Mean Square Residual using *Bentler's* formula.

The results show that measurement invariance holds for the A*dapted Thoughts' Subscale* across timepoint because at the key stages of constrained model to metric model to scalar model, the *ΔCFI* is no more than < 0.01(Chen, 2007).

Measurement Invariance for the Children's Test Anxiety Adapted Thoughts' Subscale - Group

Table G.24

Measurement Invariance of Adapted Thoughts' Subscale for Intervention and Control Groups (N = 340)

Model	Equal	X^2	*df*	*p*	*RMSEA*	*CFI*	*ΔCFI*	*TLI*	*SRMR*
Configural	Patterns	67.31	40	.00	0.07	0.96		0.95	0.04
Constrained	All	128.36	63	.00	0.09	0.91		0.92	0.08
Metric	Loadings	112.25	56	.00	0.09	0.92	0.01	0.92	0.06
Scalar	Intercepts	112.75	56	.00	0.09	0.92	0.00	0.92	0.08
Strict	Factor	127.68	62	.00	0.09	0.91	0.00	0.92	0.08
Strict	Error	98.04	55	.00	0.07	0.94	0.02	0.94	0.08

Note. ChiSq= chi-square value, d f= Degrees of Freedom; RMSEA = root mean square error of approximation; CFI comparative fit indices; TLI = Tucker-Lewis Index; SRMR=Standardised Root Mean Square Residual using *Bentler's* formula.

The results show that measurement invariance holds across the intervention and control groups at the key stages of constrained to metric to scalar as *ΔCFI* is no more than 0.01, (Chen, 2007) between theses stages.

Measurement Invariance for the Children's Test Anxiety Adapted Thoughts' Subscale – Gender

Table G.25

Measurement Invariance of Adapted Thoughts' Subscale Across Gender Groups (N = 340)

Model	Equal	X^2	*df*	*p*	*RMSEA*	*CFI*	*ΔCFI*	*TLI*	*SRMR*
Configural	Patterns	70.18	40	.00	0.08	0.95		0.95	0.05
Constrained	All	111.08	63	.00	0.08	0.93		0.94	0.11
Metric	Loadings	94.50	56	.00	0.07	0.94	0.01	0.94	0.07
Scalar	Intercepts	91.21	56	.00	0.07	0.95	0.01	0.95	0.11
Strict	Factor	103.03	62	.00	0.07	0.94	0.01	0.95	0.07
Strict	Error	108.98	55	.00	0.09	0.92	0.02	0.92	0.11

Note. ChiSq = chi-square value, df = Degrees of Freedom; RMSEA = root mean square error of approximation; CFI comparative fit indices; TLI = Tucker-Lewis Index; SRMR = Standardised Root Mean Square Residual using *Bentler's* formula.

The results show that measurement invariance holds across the key stages of the constrained to metric to scalar measures for the gender groups as the *ΔCFI* is no more than < 0.01.

ANOVAs for Chi-Square Differences for Timepoint, Intervention and Control and Gender Groups for the Children's Test Anxiety *Adapted Thoughts'* Subscale.

If Pr < (*Chi-Sq*) is significant, the null hypothesis that the models are the same is rejected and the next stage of the model is calculated for comparison. Although, the tables below detail the Chi-square values, measurement invariance was deemed to hold if the *ΔCFI* value between the constrained, metric and scalar stages was 0.01 or less (Chen, 2007).

Table G.26

ANOVAs for Timepoint Measurement Invariance (N = 340)

Model	*df*	*AIC*	*BIC*	X^2	ΔX^2	Δdf	Pr (> X^2)
Configural	40	6464.8	6648.6	84.34			
Constrained	63	6473.1	6568.8	138.62	51.15	23	< .001
Metric	56	6483.3	6605.8	134.78			
Constrained	63	6473.1	6568.8	138.63	4.49	7	0.72
Metric	56	6483.3	6605.8	134.78			
Scalar	56	6477.3	6599.8	128.80	−5.97	0	note a)
Scalar	56	6477.3	6599.8	128.80			
Strict Factor	62	6475.1	6574.7	138.62	11.04	6	0.08
Strict Error	55	6463.4	6589.7	112.91			
Strict Factor	62	6475.1	6574.7	138.62	22.96	7	< .001

Note. a) No Value; *df = Degrees of Freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Pr* (> X^2) is probability of being greater than *chi-square.*

Note. ΔCFI taken for measurement invariance.

Table G.27

ANOVAs for Intervention and Control Group Measurement Invariance (N = 340)

Model	*df*	*AIC*	*BIC*	X^2	ΔX^2	Δdf	Pr (> X^2)
Configural	40	6452.8	6636.6	86.53			
Constrained	63	6479.1	6574.8	158.81	62.67	23	< .001
Metric	56	6476.4	6598.9	142.09			
Constrained	63	6479.1	6574.8	158.81	16.58	7	.02
Metric	56	6476.4	6598.9	142.09			
Scalar	56	6479.1	6601.7	144.84	2.75	0	note a)
Strict factor	62	6480.3	6579.8	158.04			
Scalar	56	6479.1	6601.7	144.84	16.47	6	.011
Strict factor	62	6480.3	6579.8	158.04			
Strict error	55	6454.5	6580.8	118.16	26.69	7	< .001

Note. a) No Value; *df = Degrees of Freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Pr* (> X^2) is probability of being greater than *chi-square.*

Note. Δ*CFI* taken for measurement invariance.

Table G.28

ANOVAs for Gender Group Measurement Invariance (N = 340)

Model	*df*	*AIC*	*BIC*	X^2	ΔX^2	Δdf	Pr (> X^2)
Configural	40	6476.4	6660.2	98.76			
Constrained	63	6475.3	6571.0	143.67	40.99	23	< .001
Constrained		6475.3	6571.0	143.67			
Metric		6472.4	6594.9	126.75	18.58	7	0.009
Metric	56	6472.4	6594.9	126.75			
Scalar	56	6486.6	6591.1	123.02	−3.73		note a)
Scalar	56	6468.6	6591.5	123.02			
Strict Factor	62	6469.4	6568.9	135.78	12.37	6	0.05
Strict Factor	62	6469.4	6568.9	135.78			
Strict Error	55	6486.2	6612.6	138.61	−1.68	7	1.0

Note. a) No Value; *df = Degrees of Freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Pr* (> X^2) is probability of being greater than *chi-square.*

Note. Δ*CFI* taken for measurement invariance.

Answering the Research Questions

Research Question 1: Does an intervention, which uses retrieval practice to teach *the testing effect*, whilst children are practising tests, reduce the *worry* or cognitive element of test anxiety in 10 -to 11- year -olds?

Table G.29

Levene's Test for Homogeneity of Variance for the Adapted Thoughts' Subscale

Timepoint	*F*	*df*1	*df*2	*p*
Pre	0.24	1	168	0.62
Post	0.64	1	168	0.42

Main analyses: Mixed 2 x 2 Factorial ANOVA (Timepoint and Group)

Table G.30

Results for Within Subjects Effects Timepoint for the Adapted Thoughts' Subscale (N = 340)

Group	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2	η^2_p
Timepoint	188.26	1	188.26	23.55	< .001	0.021	0.123
Timepoint* Group	4.03	1	4.03	0.504	.48	< .001	0.003
Residuals	1343.21	168	11.088	7.99			

Note: Type III Sum of squares

Table G.31

Results for Between Subjects Effects Group Intervention and Control (N =340)

Group	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2	η^2_p
Group	34.94	1	34.94	0.79	.38	0.004	0.005
Residuals	7419.00	168	44.16				

Note. Type III Sum of squares.

Table G.32

Group Descriptive Statistics for the ANOVA (RQ1) (N = 340)

Timepoint	Group	*N*	*M*	*SD*	*SE*
Pre	Intervention	85	16.97	4.90	0.53
Pre	Control	85	17.39	5.13	0.56
Post	Intervention	85	15.26	5.16	0.56
Post	Control	85	16.12	5.23	0.57

Latent Means Analyses

Table G.33

Latent Means' Fit Statistics of Adapted Thoughts' Subscale Across the Timepoints Pre and Post for the pre-test and post-test Intervention Groups (n = 170)

Model	X^2	*df*	*CFI*	*TLI*	*RMSEA*	*SRMR*
Equal means	87.19	64	0.95	0.94	0.07	0.10
Unequal means	82.07	63	0.96	0.95	0.06	0.08

Note. ChiSq = chi-square value, df = Degrees of Freedom; CFI comparative fit indices; TLI = Tucker-Lewis Index; RMSEA = root mean square error of approximation; SRMR = Standardised Root Mean Square Residual using *Bentler's* formula.

Table G.34

Latent Means' ANOVA for Adapted Thoughts' Subscale Across the Timepoints Pre and Post for the pre-test and post-test Intervention Groups (n = 170)

Model	*df*	*AIC*	*BIC*	X^2	ΔX^2	Δdf	Pr ($> X^2$)
Equal Means	64	3073.2	3148.5	95.55	6.09	1	.0136
Unequal Means	63	3069.9	3148.3	90.22			

Note. *Note. df = Degrees of Freedom; AIC = Akaike information criterion; BIC = Bayesian information criterion; Pr* ($> X^2$) is probability of being greater than *chi-square.*

Table G.35

Latent Means' Fit Statistics for Adapted Thoughts' Subscale Across the Timepoints Pre and Post for the pre-test and post-test Control Groups (n = 170)

Model	X^2	*df*	*CFI*	*TLI*	*RMSEA*	*SRMR*
Equal means	89.74	64	0.93	0.94	0.07	0.12
Unequal means	87.42	63	0.93	0.94	0.07	0.11

Note. Children's Test Anxiety Scale Data

Note. *ChiSq = chi-square value, df = Degrees of Freedom; CFI comparative fit indices; TLI = Tucker-Lewis Index; RMSEA = root mean square error of approximation; SRMR = Standardised Root Mean Square Residual* using *Bentler's* formula.

Table G.36

Latent Means' ANOVA for the Adapted Thoughts' Subscale Across the Timepoints Pre and Post for the pre-test and post-test Control Groups (n = 170)

Model	*df*	*AIC*	*BIC*	X^2	ΔX^2	Δdf	Pr (> X^2)
Equal Means	64	3379.6	3457.8	110.54	2.85	1	0.09
Unequal Means	63	3379.4	3457.8	108.30			

Note. *Note. df = Degrees of Freedom; AIC = Akaike information criterion; BIC = Bayesian Information Criterion; Pr* (> X^2) is probability of being greater than *chi-square*

Multilevel Linear Analysis

Table G.37

Models, Equations and Key Comparative Statistics for Multilevel Linear Model Analysis

Model	Equation	Deviance 2LL	Intraclass Correlation	ΔX^2
Step 1 Variance Components Level 2 school random effect only.	$Yij = \gamma_{00} + \mu_{0j} + \varepsilon_{ij.}$	1034.08	0.11	
Step 2a Random Intercepts with Group as Level 1 predictor for fixed effects and School as a Level 2 predictor for random effects	$Yij = \gamma_{00} + \mu_{0j} + \gamma_{10}$ Group $+ \varepsilon_{ij}$	1033.80	0.11	1*df* / .28 non-significant change to step 1
Step 2b Random Intercepts with Group and pre-test centred mean as Level 1 predictors for fixed effects and School as a Level 2 predictor for random effects	$Yij = \gamma_{00} + \mu_{0j} + \gamma_{10}$ Group $+ \gamma_{10}$ Globcentpre $+ \varepsilon_{ij}$	927.56	0.05	1*df* / 106.24 $p < 0.01$ highly significant change to step 2a

Step 1 Results

Table G.38

Step 1 Information Criteria

Information Criteria	*Value*
−2 Log Likelihood	1034.08
Schwarz's Bayesian Criterion (BIC)	1049.48

Table G.39

Step 1 Fixed Effects - Type III Tests of Fixed Effects

Source	Numerator *df*	Denominator *df*	*F*	*p*
Intercept	1	10.03	535.14	.000

Table G.40

Step 1 Estimates of Fixed Effects

Parameter	Estimate	*SE*	*df*	*t*	*p*	95% CI Lower Bound	95% CI Upper Bound
Intercept	15.35	0.66	10.03	23.13	.000	13.87	16.83

Table G.41

Step 1 Estimates of Covariance Parameters

Parameter	Estimate	*SE*	*Wald Z*	*p*	95% CI Lower Bound	95% CI Upper Bound
Residual	24.02	2.69	8.92	.000	19.28	29.91
Intercept SchID = Variance	3.00	2.06	1.45	.15	.78	11.55

Step 2a Results

Table G.42

Step 2a Information Criteria

Information Criteria	*Value*
−2 Log Likelihood	1033.80
Schwarz's Bayesian Criterion (BIC)	1054.34

Table G.43

Step 2a Fixed Effects - Type III Tests of Fixed Effects

Source	Numerator *df*	Denominator *df*	*F*	*p*
Intercept	1	10.36	285.01	.000
Group	1	9.78	0.29	.605

Table G.44

Step 2a Estimates of Fixed Effects

Parameter	Estimate	*SE*	*df*	*t*	*p*	95% CI Lower Bound	95% CI Upper Bound
Intercept	15.04	0.89	10.36	16.88	.000	13.06	17.01
Group	0.70	1.31	9.78	0.53	.605	−2.23	3.62

Table G.45

Step 2a Estimates of Covariance Parameters

Parameter	Estimate	*SE*	*Wald Z*	*p*	95% CI Lower Bound	95% CI Upper Bound
Residual	24.03	2.69	8.92	.000	19.28	29.93
Intercept SchID = Variance	2.84	2.00	1.41	.157	0.71	11.36

Step 2b

Table G.46

Step 2b Information Criteria

Information Criteria	*Value*
−2 Log Likelihood	927.56
Schwarz's Bayesian Criterion (BIC)	953.24

Table G.47

Step 2b Fixed Effects - Type III Tests of Fixed Effects

Source	Numerator *df*	Denominator *df*	*F*	*p*
Intercept	1	9.49	864.14	.000
Group	1	8.45	0.225	.647
Pre-test Centred Mean	1	169.80	149.88	.000

Table G.48

Step 2b Estimates of Fixed Effects

Parameter	Estimate	*SE*	*df*	*t*	*p*	95% CI Lower Bound	95% CI Upper Bound
Intercept	15.37	.52	9.49	29.40	.000	14.20	16.54
Group	0.36	.76	8.45	.475	.647	−1.38	2.10
Pre-test Centred Mean	0.70	.58	169.80	12.24	.00	0.59	0.82

Table G.49

Step 2b Estimates of Covariance Parameters

Parameter	Estimate	*SE*	*Wald Z*	*p*	95% CI Lower Bound	95% CI Upper Bound
Residual	13.24	1.49	8.90	.000	10.63	16.51
Intercept SchID = Variance	0.64	0.71	0.90	.367	0.73	5.58

Research Question 4 : Is there a significant difference between the groups in the intervention group where the researcher delivered the intervention?

Table G.50

Levene's Test for Homogeneity of Variance for the Experimenter Effect / Fidelity Treatment; researcher groups vs teacher groups

Timepoint	*F*	*df*1	*df*2	*p*
Pre	0.20	1	60	.65
Post	0.03	1	60	.87

Mixed 2 x 2 Factorial ANOVA For Experimenter Bias, Fidelity.

Table G.51

Results for Within Subjects Effects for Timepoint for the Researcher Vs Teacher Groups Using the Adapted Thoughts' Subscale to Measure Test Anxiety (N = 62)

Cases	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2	η^2_p
Timepoint	105.03	1	105.03	3.52	.07	0.014	0.06
Timepoint* Group	5.16	1	5.16	0.17	.68	< .001	0.003
Residuals	1789.68	60	29.83				

Note. Type III Sum of squares.

Table G.52

Results for Between Subjects Effects for Timepoint for the Researcher Vs Teacher Groups Using the Adapted Thoughts' Subscale to Measure Test Anxiety (N = 62)

Cases	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2	η^2_p
ID	0.32	1	0.32	0.003	.954	4.17 e−5	5.53 e−5
Residuals	5835.62	60	97.26				

Note. Type III Sum of squares.

Table G.53

Descriptive Statistics for Research Question 4 (N = 62)

Timepoint	ID	*Mean*	*SD*	*N*
Level 1 (Post-test)	Researcher	27.66	9.11	32
	Teacher	28.17	8.07	30
Level 2 (Pre-test)	Researcher	29.91	7.17	32
	Teacher	29.60	7.36	30

Research Question 5: The Covid SARS – COV 2 Question. Is There a Difference in Feelings of Test Anxiety Between Students Who Sit High Stakes' Paper Tests or Online Tests?

Table G.54

Levene's Test for Homogeneity of Variance for Paper Vs Online High Stakes' Tests

Timepoint	*F*	*df*1	*p*
Pre	0.60	1	.44
Post	1.62	1	.21

Table G.55

Independent Samples T Test – Test Anxiety with High Stakes' Paper Tests vs Online Tests (N = 51)

Timepoint	*t*	*df*	*p*	Cohen's *d*
Pre	.60	49	.55	0.17
Post	.38	49	.71	0.11

Note. Student's t-test.

Table G.56

Descriptive Statistics for Independent T Test – Test Anxiety with High Stakes' Tests vs Online Tests (N =51)

Timepoint	Group	*N*	*Mean*	*SD*	*SE*
Pre	online	30	17.83	5.43	0.99
	paper	21	16.95	4.72	1.03
Post	online	30	17.37	5.60	1.02
	paper	21	16.81	4.59	1.00

Research Question 2b

Table G.57

Multi Indicators Multiple Causes (MIMIC Models)

Model and Covariate	Estimate	*SE*	*z*	*p*
1.Intervention pre-test with SETT pre-test	-0.04	0.02	-2.54	.01
2. Intervention pre-test with LSETT pre-test	0.07	0.02	2.76	.006
3. Control pre-test with SETT pre-test	-0.001	0.002	-4.08	.68
4. Control pre-test with LSETT pre-test	0.002	0.006	0.39	.69
5. Intervention post-test with SETT post-test	-0.04	0.02	-2.09	.04
6. Intervention post-test with LSETT post-test	0.12	0.04	3.25	.001
7. Control post-test with SETT post-test	-0.01	0.023	1.19	.29
8. Control post-test with LSETT post-test	0.022	0.05	0.43	.24
9. Intervention post-test with SETT pre-test	-0.04	0.02	-2.12	.03
10. Intervention post-test with LSETT pre-test	0.10	0.03	3.24	.001
11. Control post-test with SETT pre-test	0.008	0.01	0.58	.56
12. Control post-test with LSETT pre-test	0.01	0.02	0.52	.10

Figure G.17

Intervention, Pre-test with Self-efficacy in Test Taking Sum Score as Covariate

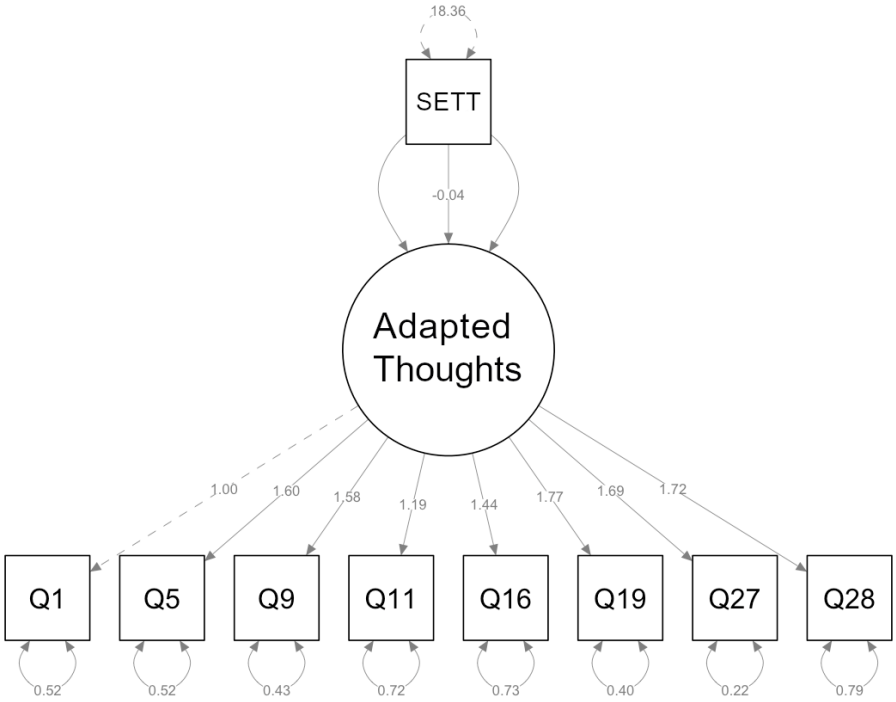


Figure G.18

Intervention Pre-test with Little Self-efficacy in Test Taking Sum Score as Covariate

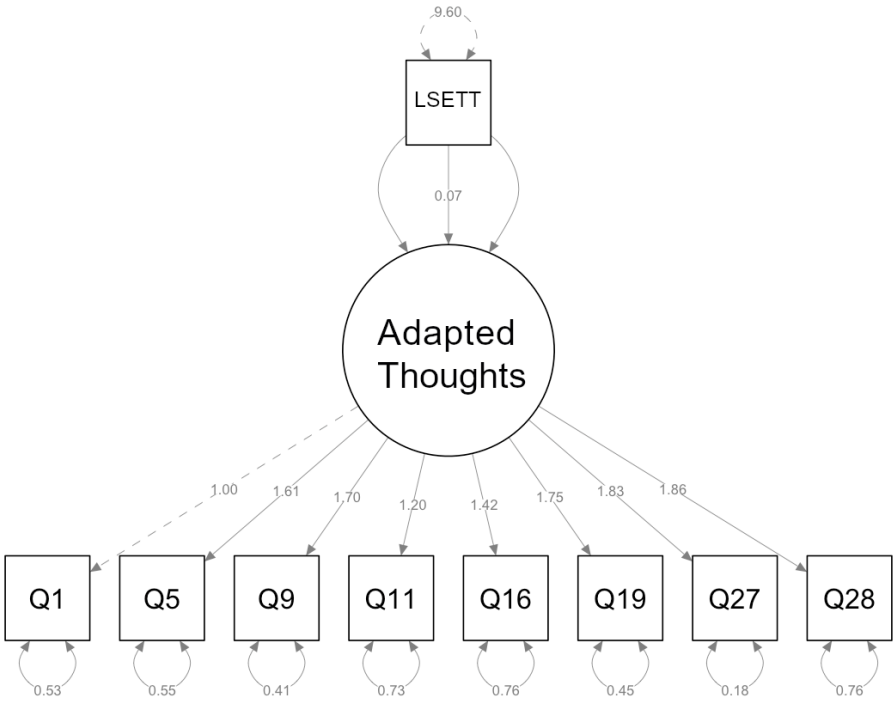


Figure G.19

Control, Pre-test with Self-efficacy in Test Taking Sum Score as Covariate

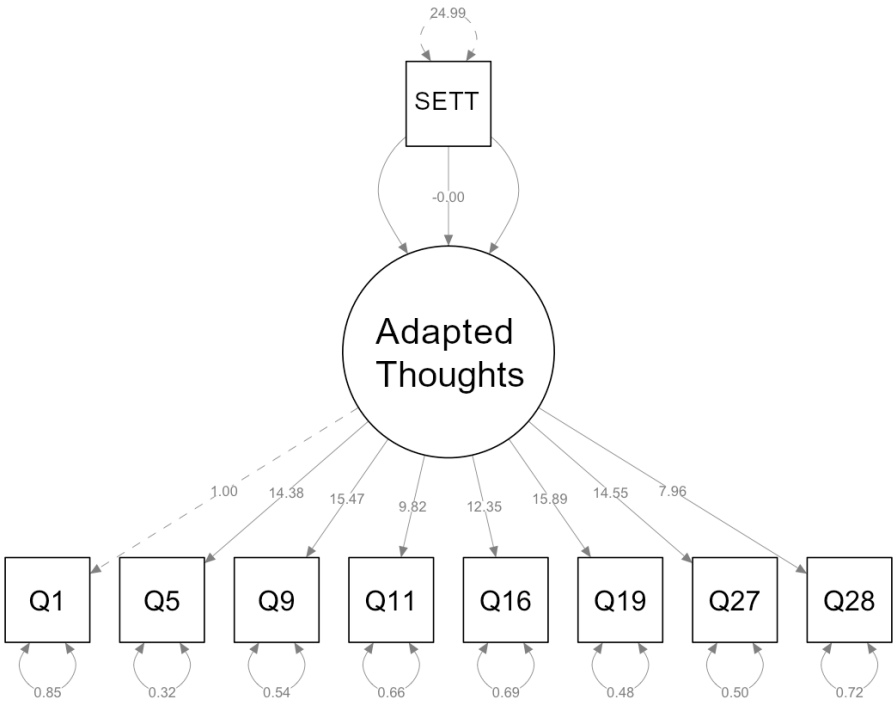


Figure G.20

Control, Pre-test with Little Self-efficacy in Test Taking Sum Score as Covariate

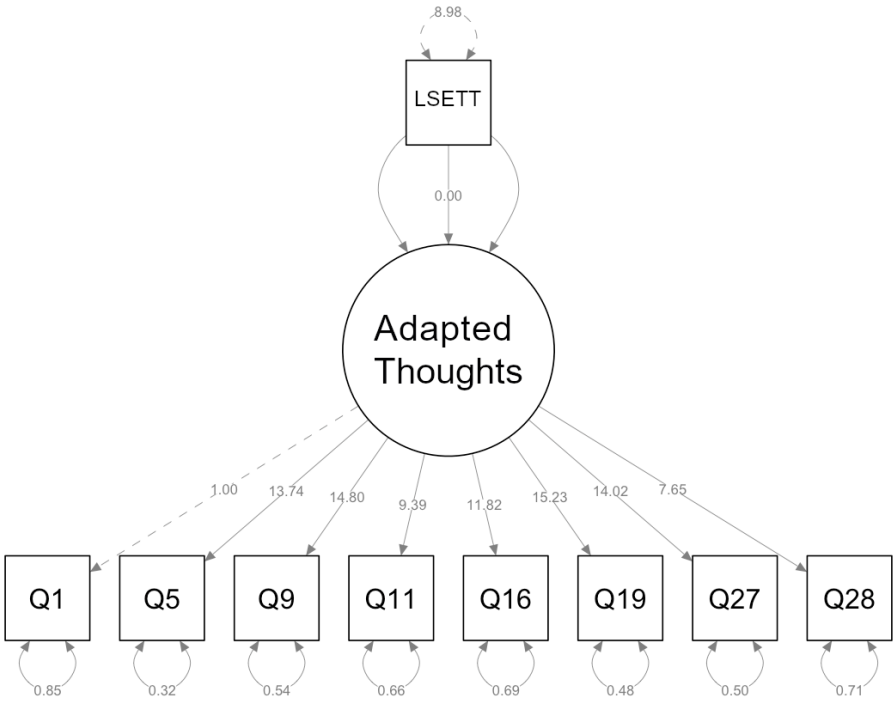


Figure G.21

Intervention post-test with Self-efficacy in Test Taking Post-test Sum Score as Covariate

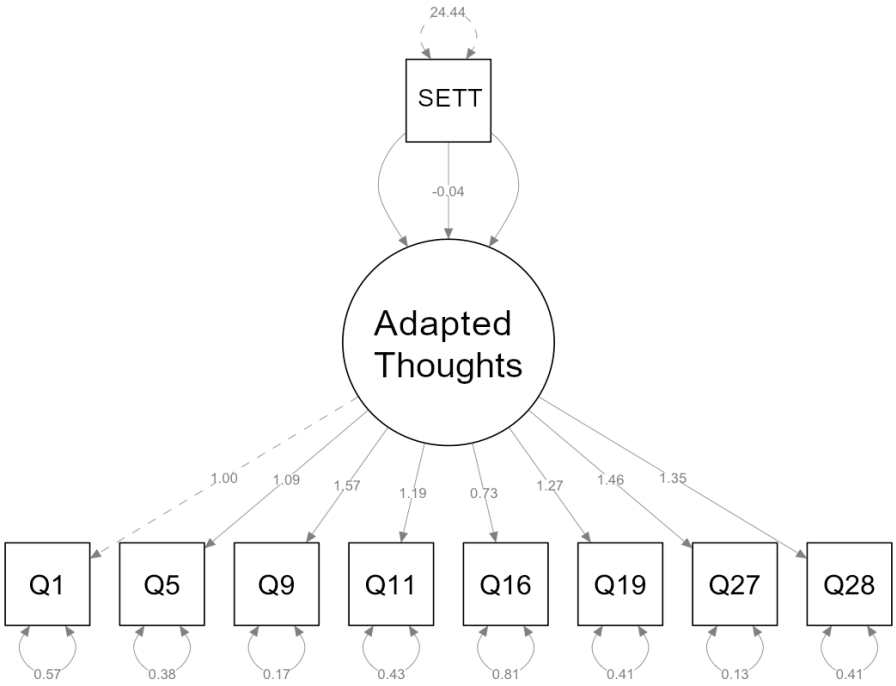


Figure G.22

Intervention post-test, Little Self-efficacy in Test Taking Post-test Sum Score as Covariate

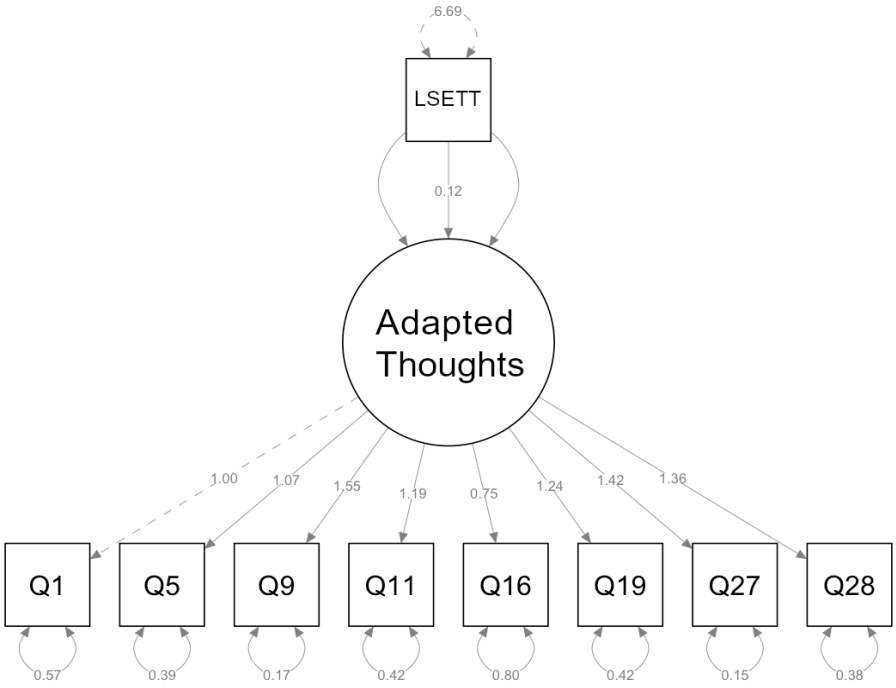


Figure G.23

Control, Post-test with Self-efficacy in Test Taking Sum Score as Covariate

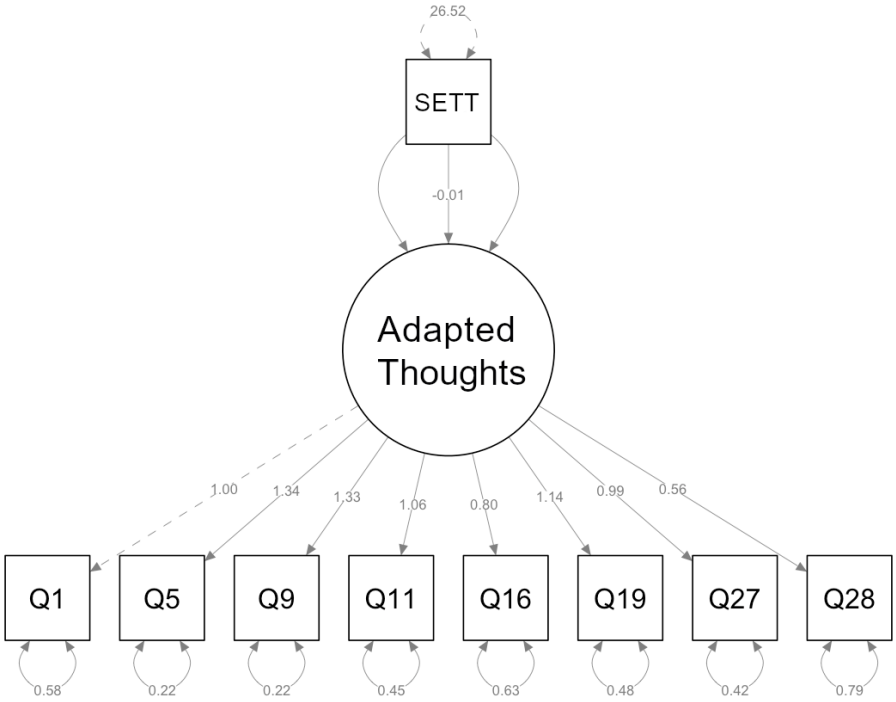


Figure G.24

Control, Post-test with Little Self-efficacy in Test Taking Sum Score as Covariate

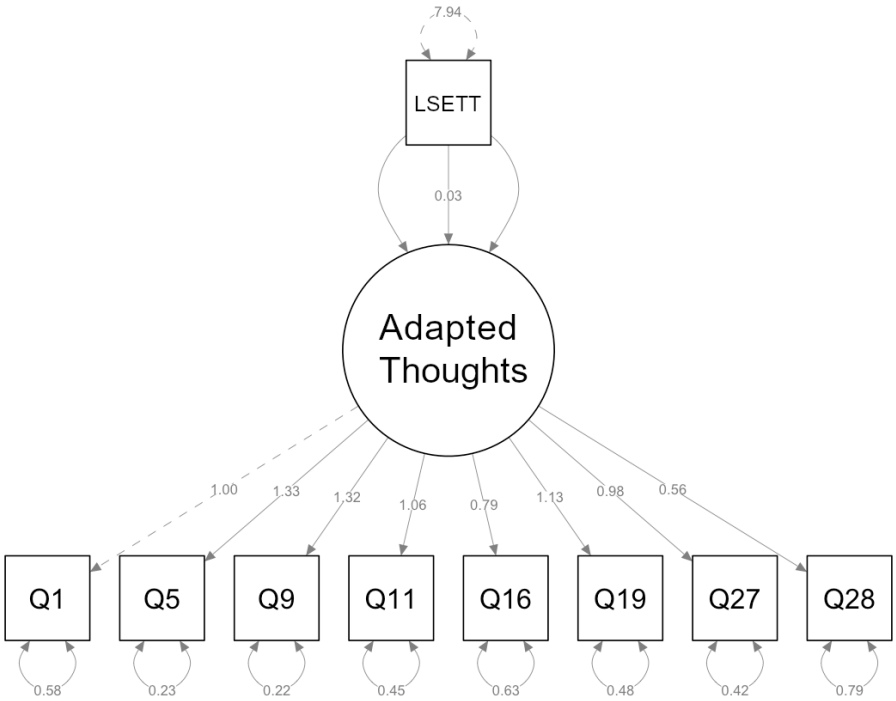


Figure G.25

Intervention post-test with Self-efficacy in Test Taking Pre-test Sum Score as Covariate

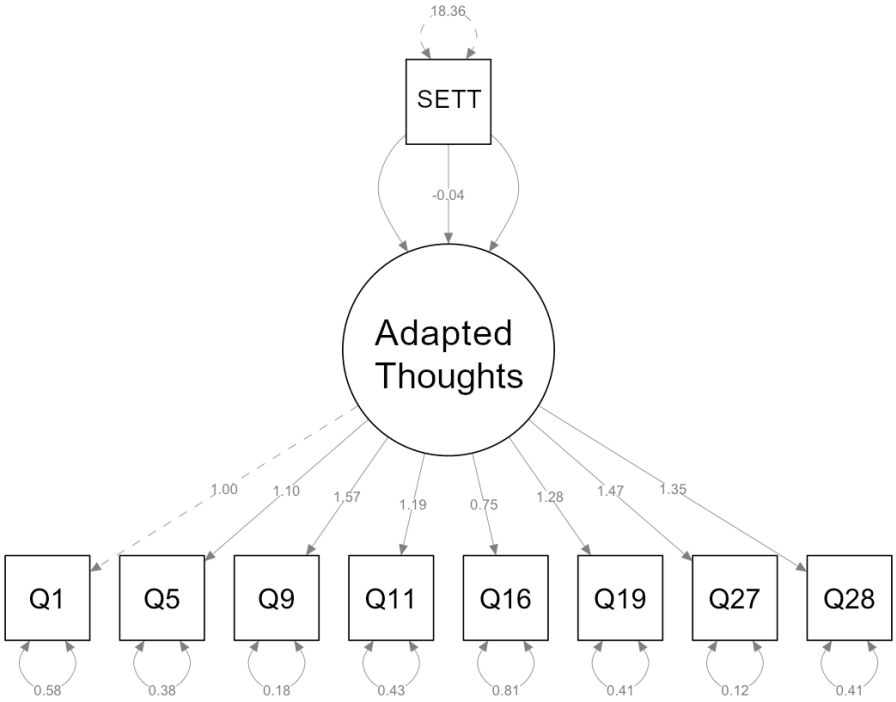


Figure G.26

Intervention post-test with Little Self-efficacy in Test Taking Pre-test Sum Score as Covariate

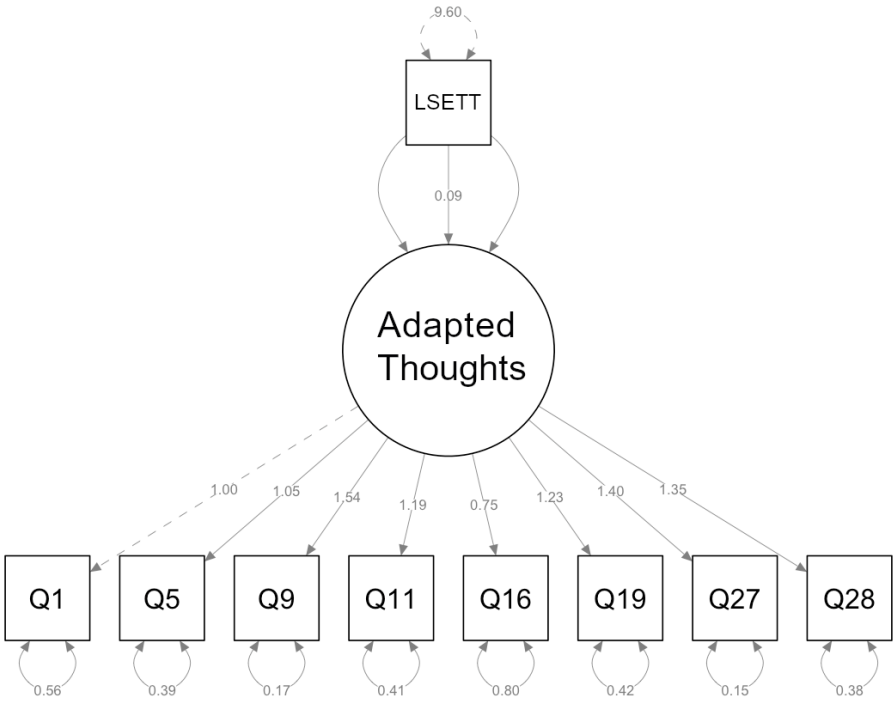


Figure G.27

Control post-test with Self-efficacy in Test Taking Pre-test Sum Score as Covariate

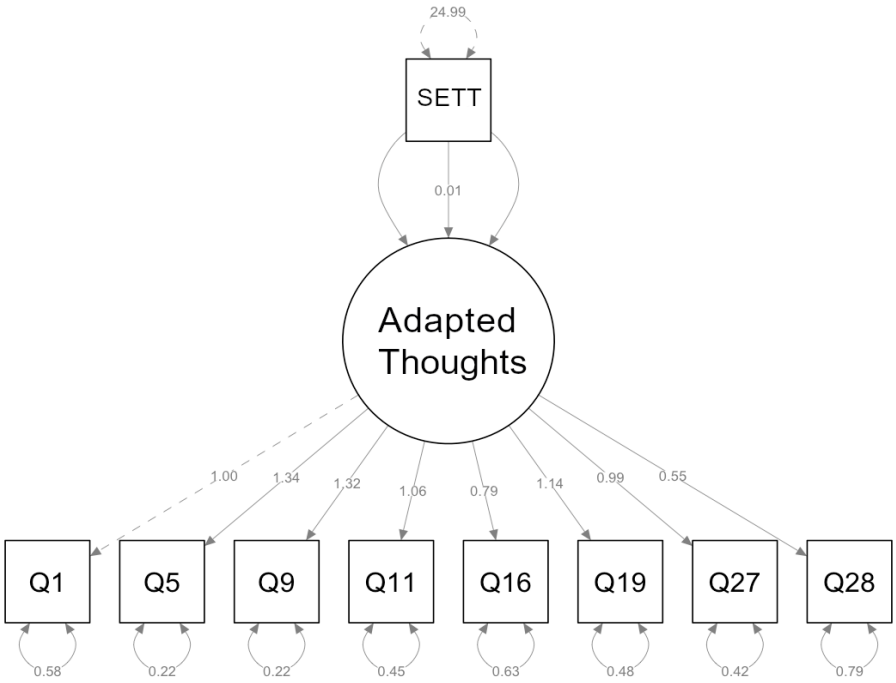
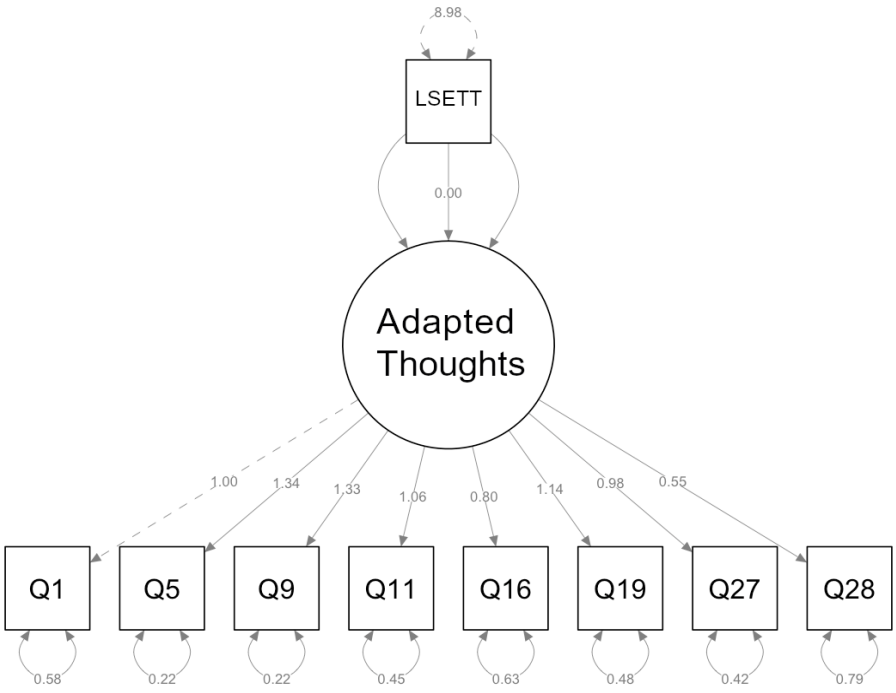


Figure G.28

Control post-test with Little Self-efficacy in Test Taking Pre-test Sum Score as Covariate



Appendix H

Results of the Self-efficacy in Test Taking Instrument

Descriptive Statistics

Table H.1

Means, Standard Deviations, Skew and Kurtosis for the Self-efficacy in Test Taking Instrument with the Pre-test, Intervention Group (n = 59)

Question	*M*	*SD*	Skew	Kurtosis	*SE*
Q1 test hard - keep try	3.15	0.89	−0.44	−1.26	0.12
Q2 test hard - confident	2.58	0.83	0.20	−0.72	0.11
Q3 badly - positive	2.32	0.82	0.29	−0.44	0.11
Q4 test hard - not pass	2.27	0.96	0.37	−0.84	0.13
Q5 take test - anxious instructor	1.69	1.04	1.16	−0.10	0.14
Q6 take test - anxious room	1.69	0.84	0.95	0.02	0.11
Q7 test hard - negative thought	2.31	1.05	0.08	−1.31	0.14
Q8 test hard - doubt pass	2.20	0.98	0.25	−1.06	0.13
Q9 test-panic if others finish	2.63	1.19	−0.12	−1.54	0.15
Q10 concentrate -even difficult	2.63	0.93	0.02	−0.97	0.12
Q11 enjoy tests	2.29	1.08	0.30	−1.22	0.14
Q12 cheerful when test hard	2.98	0.92	−0.36	−1.01	0.12
Q13 manage even when unwell	2.34	0.98	0.28	−0.94	0.13
Q14 test difficult -not end of world	3.15	0.89	−0.73	−0.40	0.12
Q15 do not worry about tests	2.37	1.00	0.24	−1.03	0.13

Table H.2

Means, Standard Deviations, Skew and Kurtosis for Self-efficacy in Test Taking Instrument with the Post-test, Intervention Group (n = 59)

Question	*M*	*SD*	Skew	Kurtosis	*SE*
Q1 test hard - keep try	3.34	0.76	−0.63	−1.03	0.10
Q2 test hard - confident	2.71	0.83	0.03	−0.81	0.11
Q3 badly - positive	2.51	0.97	0.09	−1.03	0.13
Q4 test hard - not pass	2.00	0.87	0.62	−0.29	0.11
Q5 take test - anxious instructor	1.39	0.72	2.02	3.89	0.09
Q6 take test - anxious room	1.41	0.65	1.66	2.95	0.08
Q7 test hard - negative thought	1.95	0.86	0.90	0.39	0.11
Q8 test hard - doubt pass	2.00	0.85	0.66	−0.09	0.11
Q9 test-panic if others finish	2.14	1.07	0.47	−1.09	0.14
Q10 concentrate -even difficult	2.75	1.03	−0.24	−1.15	0.13
Q11 enjoy tests	2.27	1.03	0.38	−1.02	0.13
Q12 cheerful when test hard	3.14	0.84	−0.59	−0.52	0.11
Q13 manage even when unwell	2.42	1.04	0.02	−1.22	0.14
Q14 test difficult -not end of world	3.25	0.88	−0.95	−0.03	0.11
Q15 do not worry about tests	2.41	1.02	0.39	−0.66	0.13

Table H.3

Means, Standard Deviations, Skew and Kurtosis for Self-efficacy in Test Taking Instrument with the Pre-test, Control Group (n = 86)

Question	*M*	*SD*	Skew	Kurtosis	*SE*
Q1 test hard - keep try	3.25	0.76	−0.68	−0.23	0.10
Q2 test hard - confident	2.37	0.85	0.56	−0.39	0.11
Q3 badly - positive	2.29	0.85	0.09	−0.75	0.11
Q4 test hard - not pass	2.07	0.87	0.65	−0.14	0.11
Q5 take test - anxious instructor	1.71	0.95	1.18	0.34	0.12
Q6 take test - anxious room	1.41	0.67	1.67	2.59	0.09
Q7 test hard - negative thought	2.05	0.99	0.63	−0.66	0.13
Q8 test hard - doubt pass	2.15	0.87	0.49	−0.38	0.11
Q9 test-panic if others finish	2.05	0.94	0.40	−0.92	0.12
Q10 concentrate - even difficult	2.78	0.85	−0.07	−0.87	0.11
Q11 enjoy tests	1.85	0.98	0.95	−0.17	0.13
Q12 cheerful when test hard	3.00	0.91	−0.27	−1.22	0.12
Q13 manage even when unwell	2.49	0.92	0.22	−0.86	0.12
Q14 test difficult -not end of world	3.20	1.01	−0.99	−0.29	0.13
Q15 do not worry about tests	2.68	1.01	−0.13	−1.14	0.13

Table H.4

Means, Standard Deviations, Skew and Kurtosis for Self-efficacy in Test Taking Instrument with the Post-test, Control Group (n = 86)

Question	*M*	*SD*	Skew	Kurtosis	*SE*
Q1 test hard - keep try	3.45	0.82	−1.36	0.95	0.09
Q2 test hard - confident	2.66	0.94	0.12	−1.10	0.10
Q3 badly - positive	2.67	0.96	−0.03	−1.08	0.10
Q4 test hard - not pass	1.93	0.79	0.82	0.58	0.09
Q5 take test - anxious instructor	1.56	0.94	1.64	1.51	0.10
Q6 take test - anxious room	1.55	0.82	1.49	1.54	0.09
Q7 test hard - negative thought	2.02	1.06	0.71	−0.77	0.11
Q8 test hard - doubt pass	2.03	0.86	0.70	0.00	0.09
Q9 test-panic if others finish	2.06	1.10	0.57	−1.07	0.12
Q10 concentrate - even difficult	2.88	0.91	−0.23	−1.00	0.10
Q11 enjoy tests	1.78	0.94	1.12	0.35	0.10
Q12 cheerful when test hard	3.16	0.93	−0.58	−1.04	0.10
Q13 manage even when unwell	2.64	1.02	−0.04	−1.18	0.11
Q14 test difficult -not end of world	3.29	0.93	−0.94	−0.42	0.10
Q15 do not worry about tests	2.48	1.16	−1.47	−1.47	0.12

Figure H.1

Individual Histograms of Each Item on the Self-efficacy in Test Taking Questionnaire -Pre

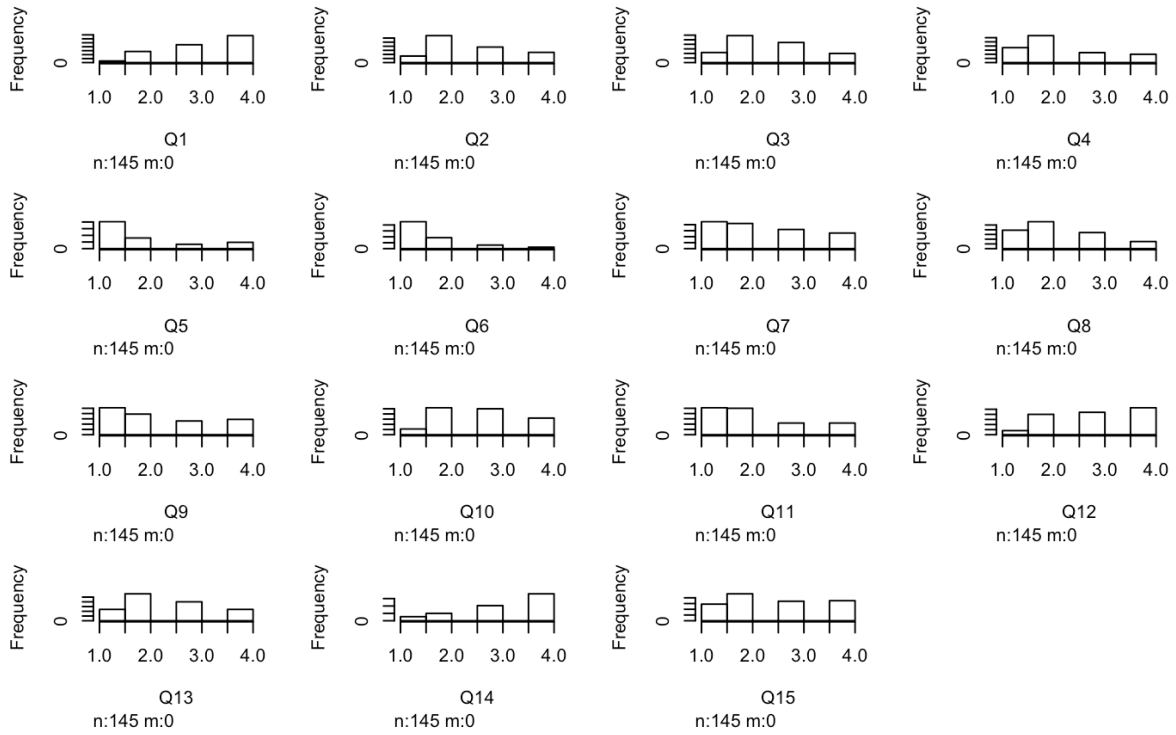
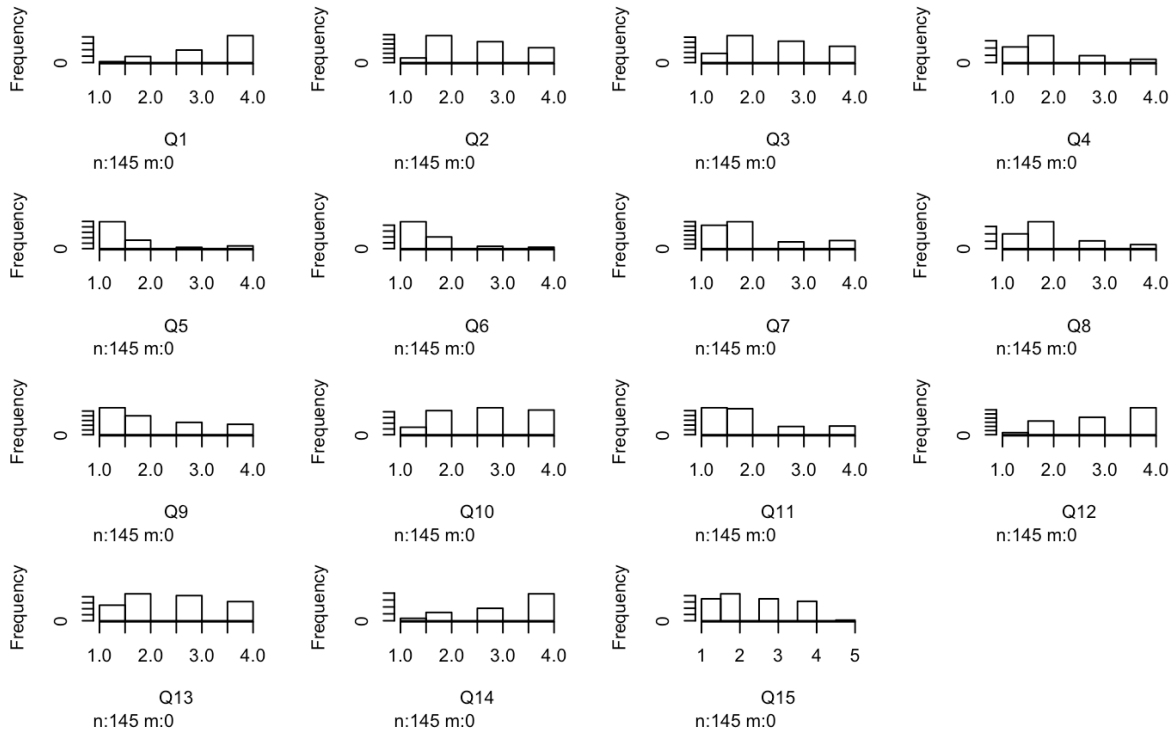


Figure H.2

Individual Histograms of Each Item on the Self-efficacy in Test Taking Questionnaire -Post



Assumptions

Violin Plots: Response Frequencies for Self-efficacy in Test Taking Instrument Pre-test and Post-test, Intervention and Control Groups – Items Q5 and Q6 Removed.

Figure H.3

Self-efficacy in Test Taking Intervention Group Pre-test, n = 59

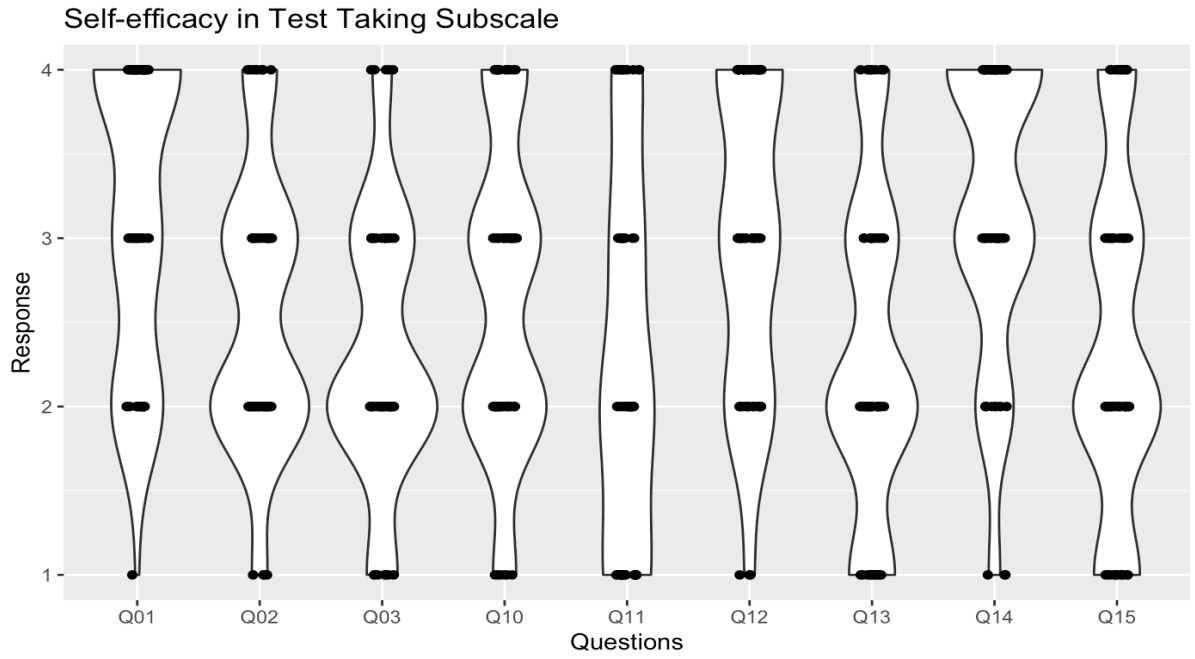


Figure H.4

Self-efficacy in Test Taking Intervention Group Post-test, n = 59

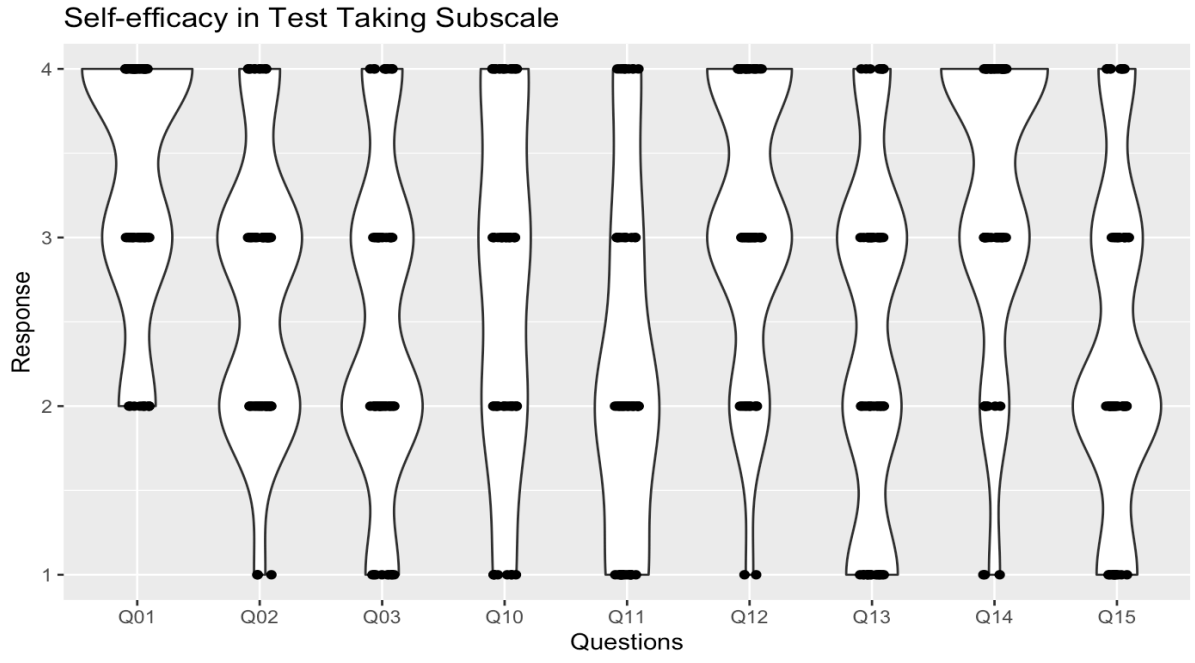


Figure H.5

Self-efficacy in Test Taking Control Group Pre-test, n = 86

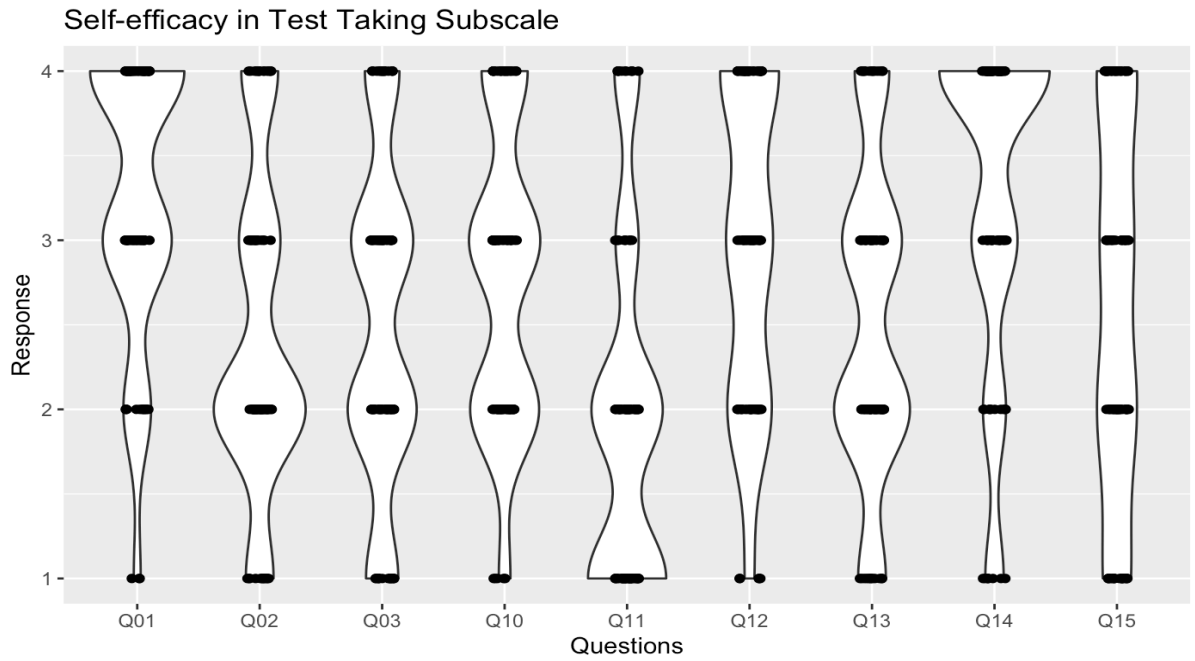


Figure H.6

Self-efficacy in Test Taking Control Group Post-test, n = 86

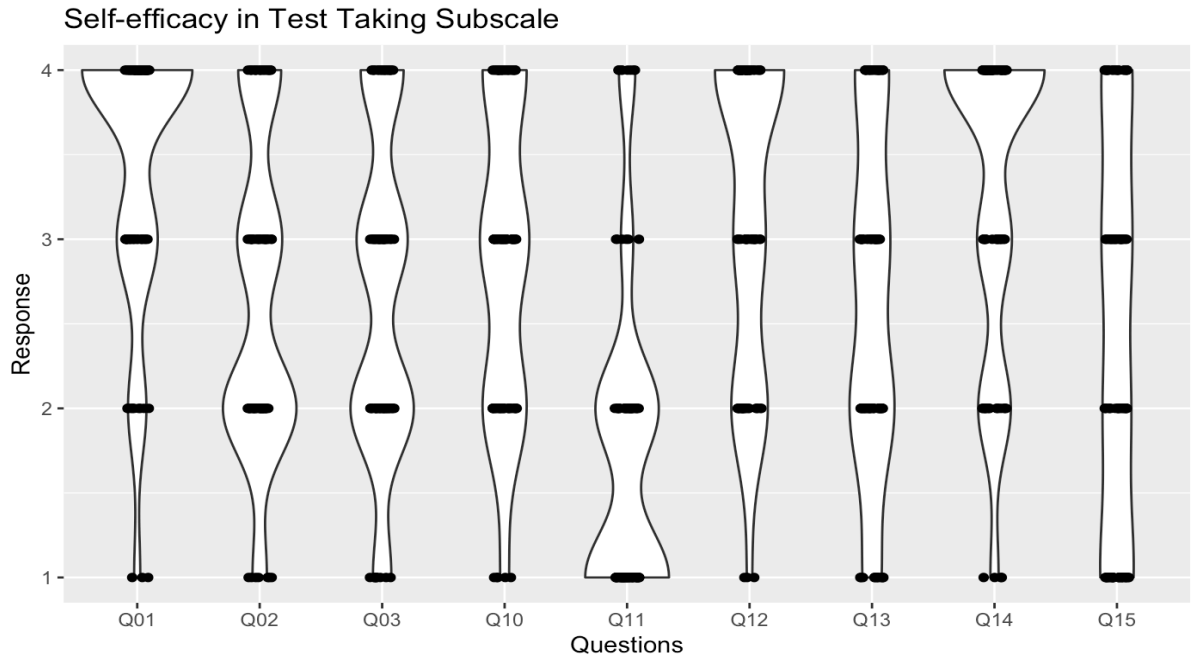


Figure H.7

Little Self-efficacy in Test Taking Intervention Group Pre-test, n = 59

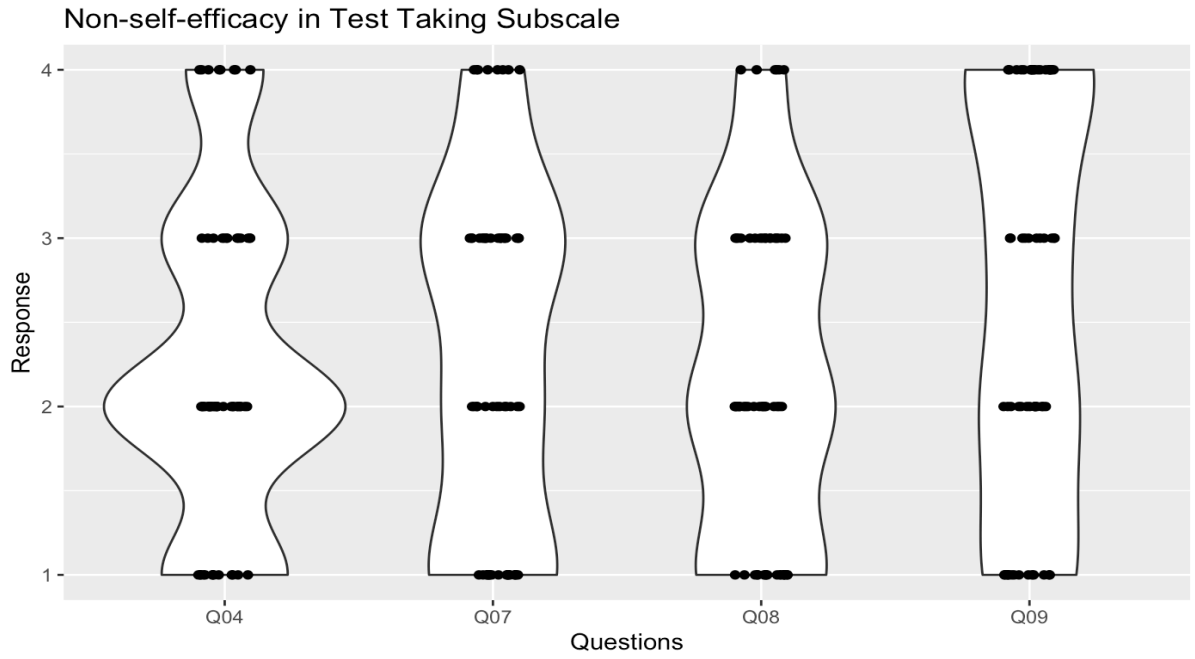


Figure H.8

Little Self-efficacy in Test Taking Intervention Group Post-test, n = 59

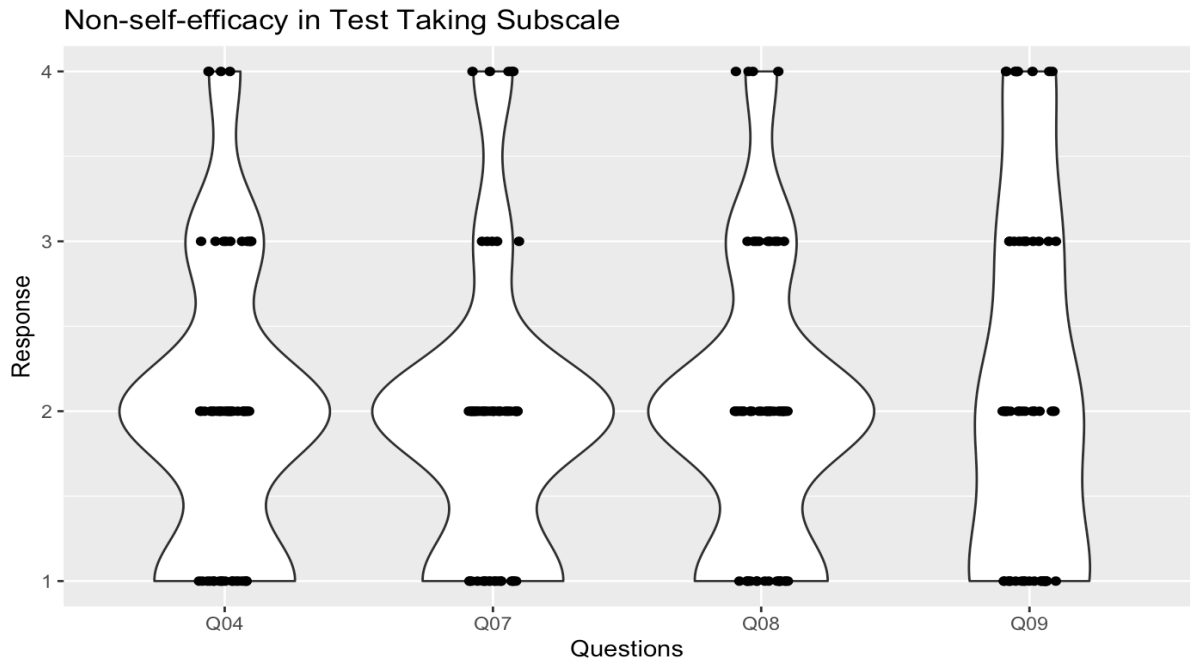


Figure H.9

Little Self-efficacy in Test Taking Control Group Pre-test, n = 86

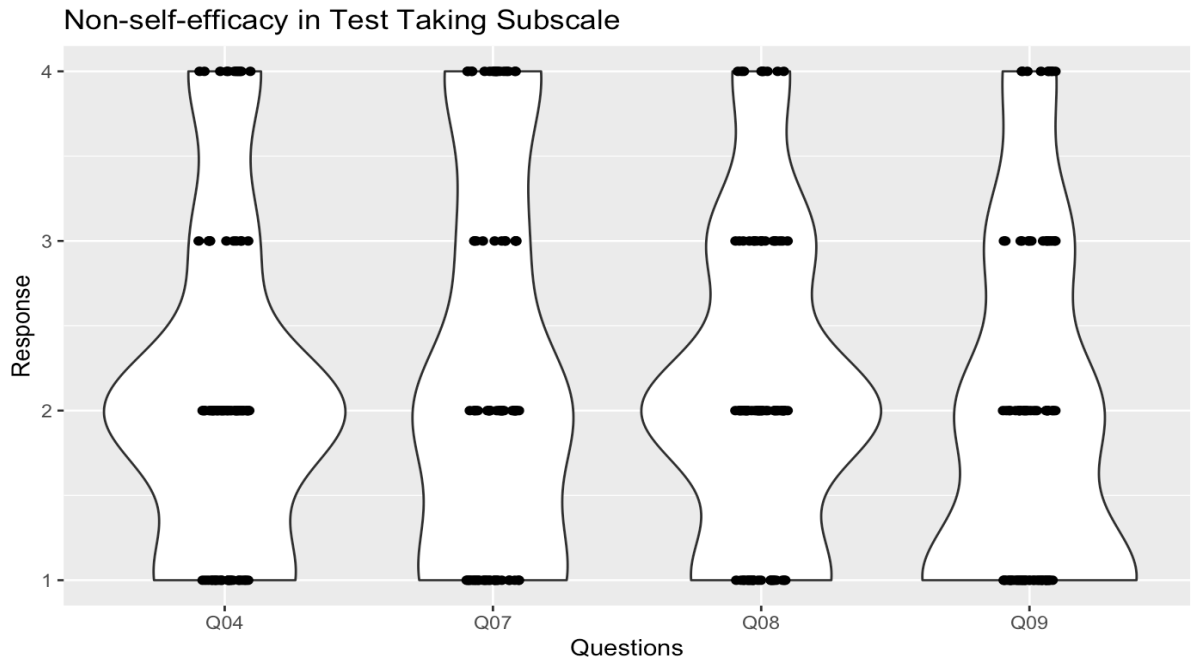
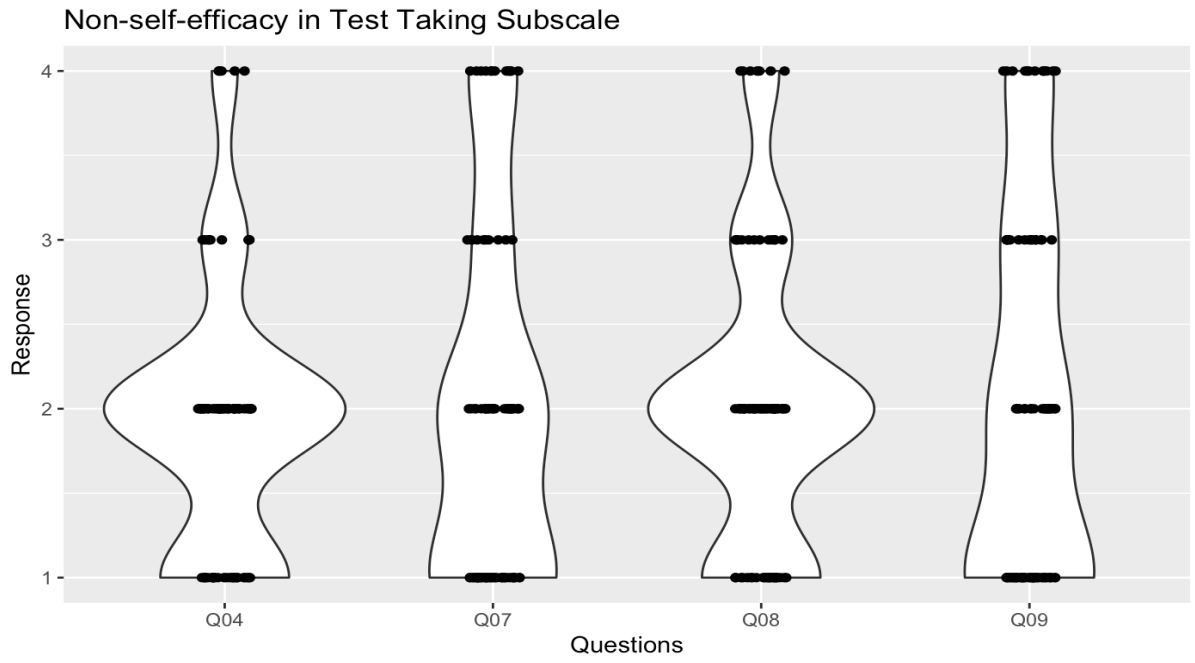


Figure H.10

Little Self-efficacy in Test Taking Control Group Post-test, n = 86



Exploratory Factor Analysis

Assumptions

Table H.5

Shapiro-Wilk Test for Normality of Distribution for Pre and Post Timepoints and Intervention and Control Group with the Self-efficacy Data (*N* = 290)

Timepoint	Group	*Shapiro-Wilk*	
		Statistic	*p*
Pre-test	intervention	0.87	< .001
	control	0.86	< .001
Post-test	intervention	0.87	< .001
	control	0.85	< .001

Table H.6

Kaiser-Meyer-Olkin Measures of Sampling Adequacy for the Self-efficacy Instrument (*N* = 290)

Timepoint	Self-Efficacy	
	Intervention	Control
Pre	0.54	0.80
Post	0.70	0.78

Table H.7

Bartlett's Test for Sphericity for the Self-efficacy Instrument (*N* = 290)

Timepoint	Group	Instrument	X^2	*df*	*p*
Pre-test	Intervention	Self-efficacy	262.88	105	< .001
	Control	Self-efficacy	341.94	105	< .001
Post-test	Intervention	Self-efficacy	282.49	105	< .001
	Control	Self-efficacy	390.25	105	< .001

Correlations

Table H.8

Determinant Matrix Values for the Self-efficacy in Test Taking at Pre-test and Post-test with the Intervention and Control Groups. (*N* = 290)

Timepoint	Group	*R*
Pre-test	Intervention	0.00647
	Control	0.01331
Post-test	Intervention	0.00432
	Control	0.00723

Confirmatory Factor Analysis

Table H.9

Shapiro-Wilk Test for Normality of Distribution for Pre and Post Timepoints and Intervention and Control Group for the Self-efficacy in Test Taking Instrument ($N = 290$)

Timepoint	Group	*Shapiro-Wilk*	
		Statistic	*p*
Pre-test	Intervention	0.87	< .001
	Control	0.87	< .001
Post-test	Intervention	0.87	< .001
	Control	0.85	< .001

Correlation Matrices for the Self-efficacy in Test Taking Instrument, X = Significant Value

Figure H.11

Self-efficacy in Test Taking Intervention Pre-test, n = 59

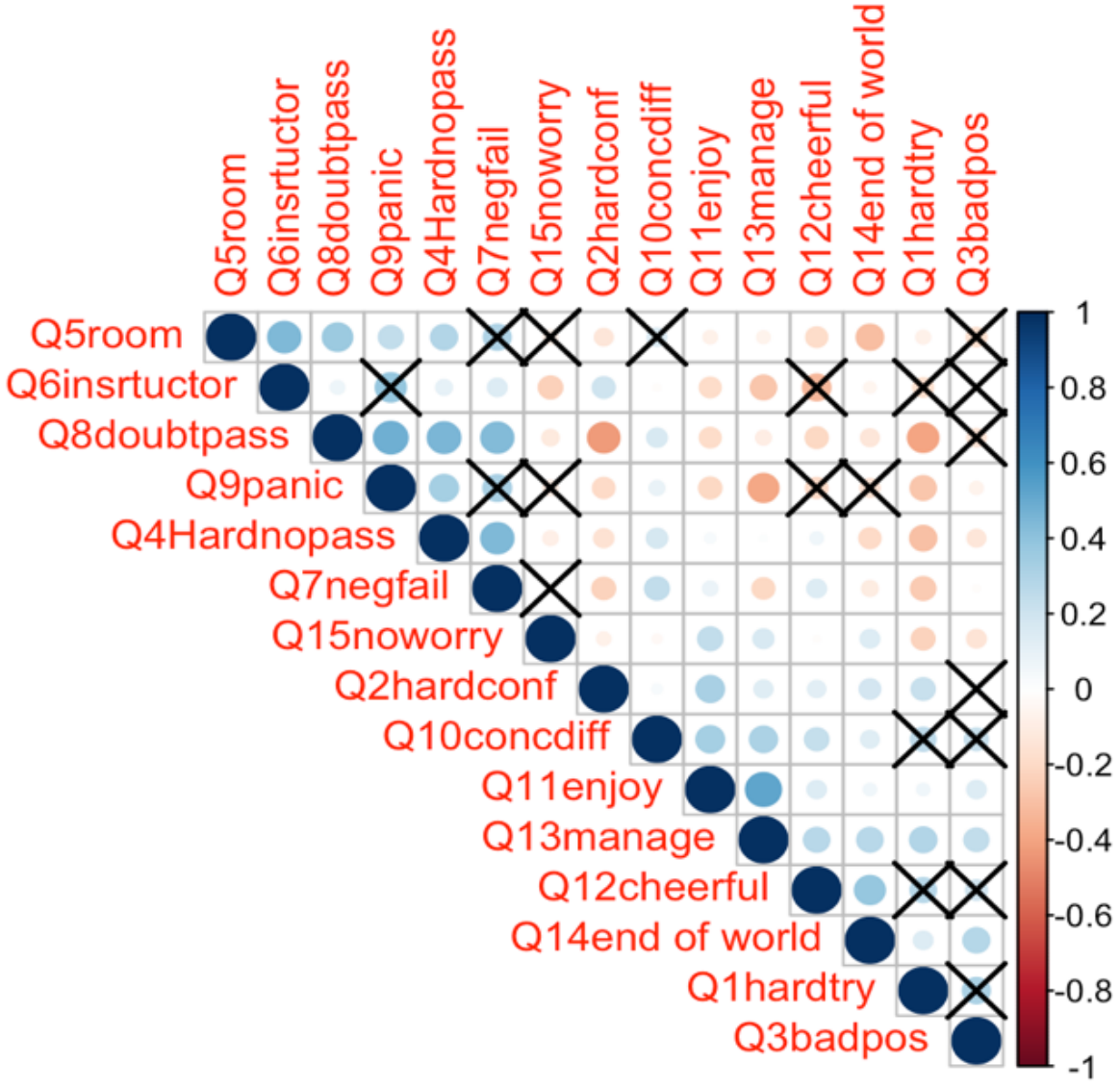


Figure H.12

Self-efficacy in Test Taking Intervention Post-test, n = 59

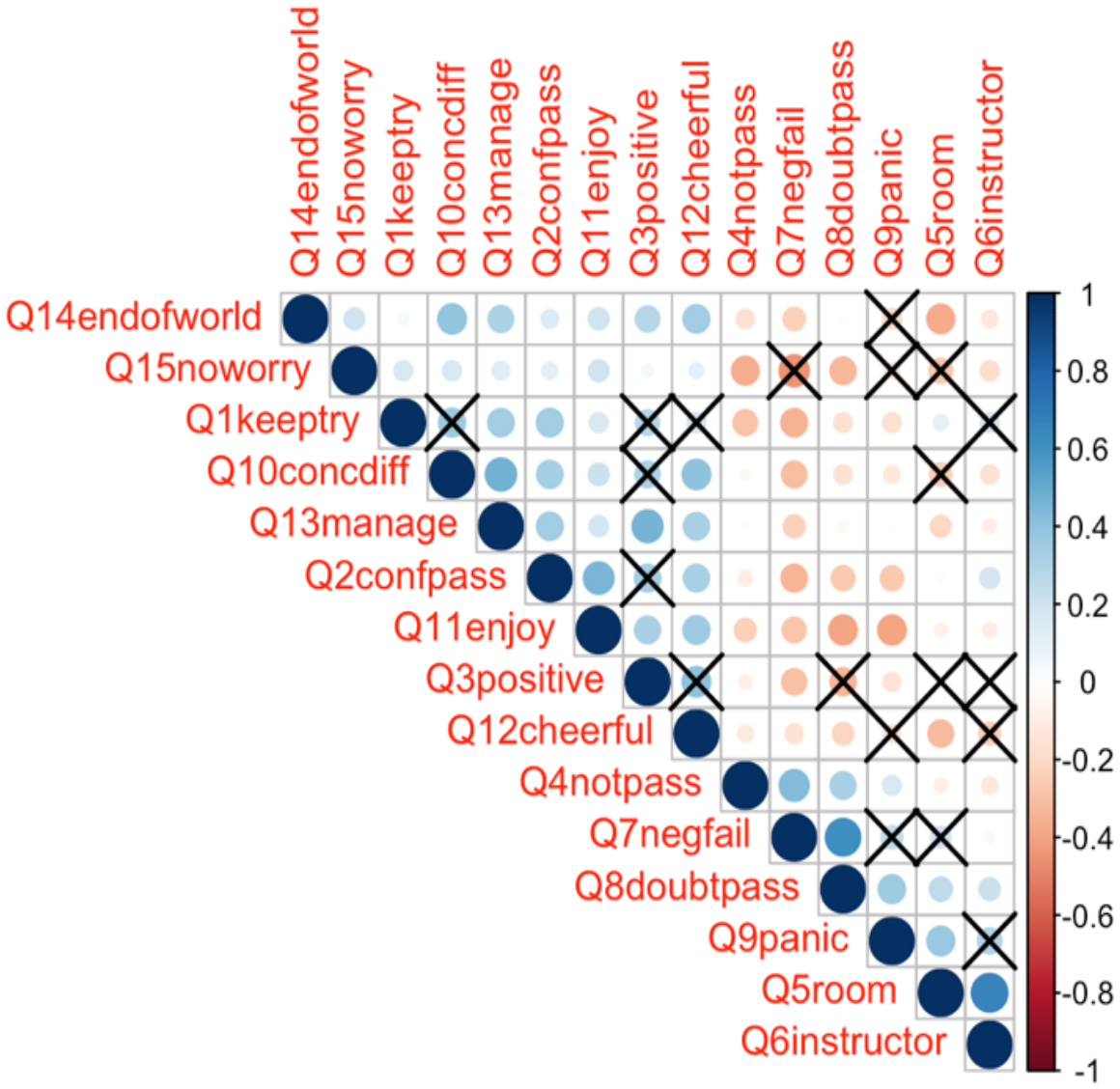


Figure H.13

Self-efficacy in Test Taking Control Pre-test, n = 86

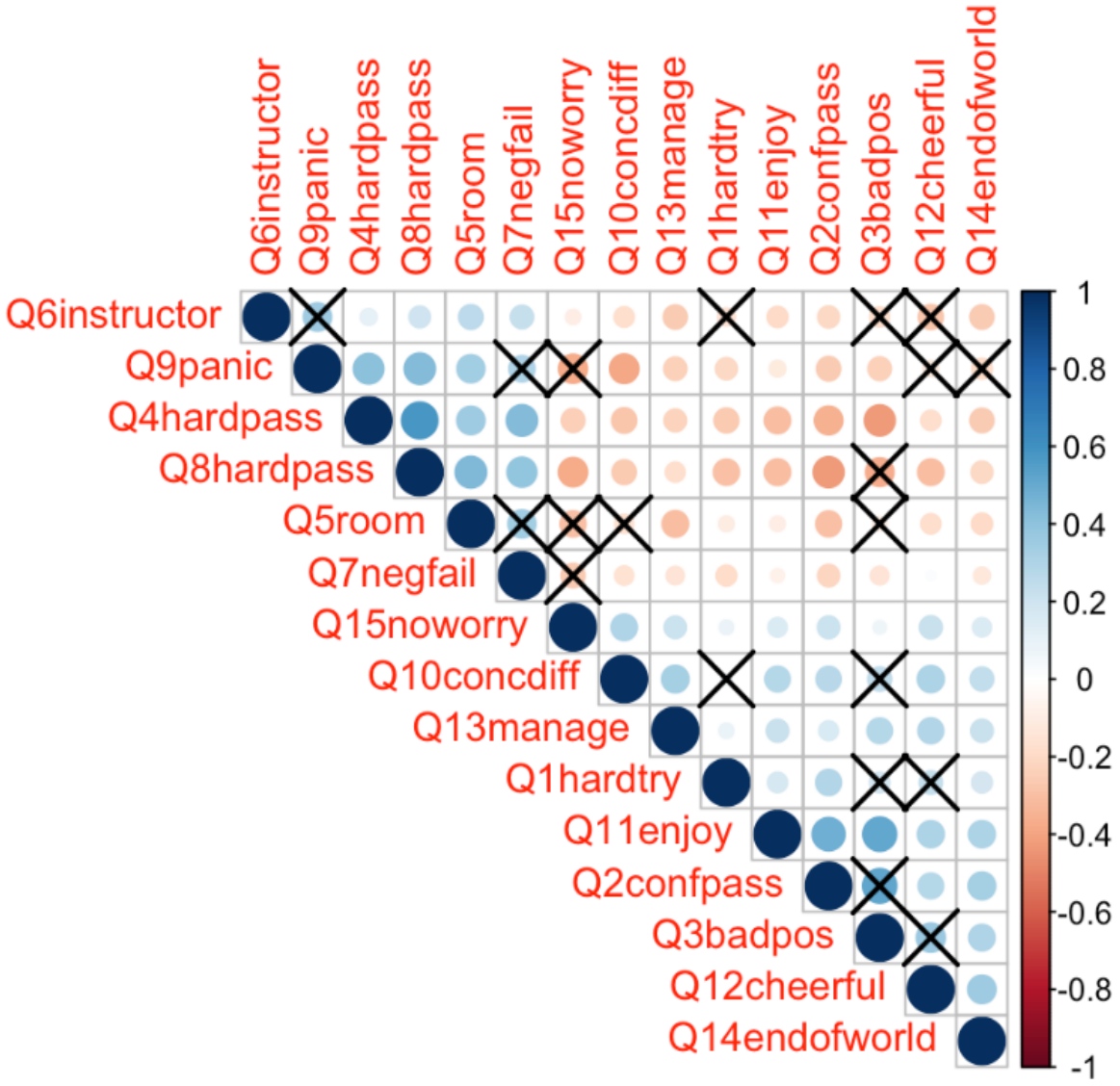


Figure H.14

Self-efficacy in Test Taking Control Post-test, n = 86

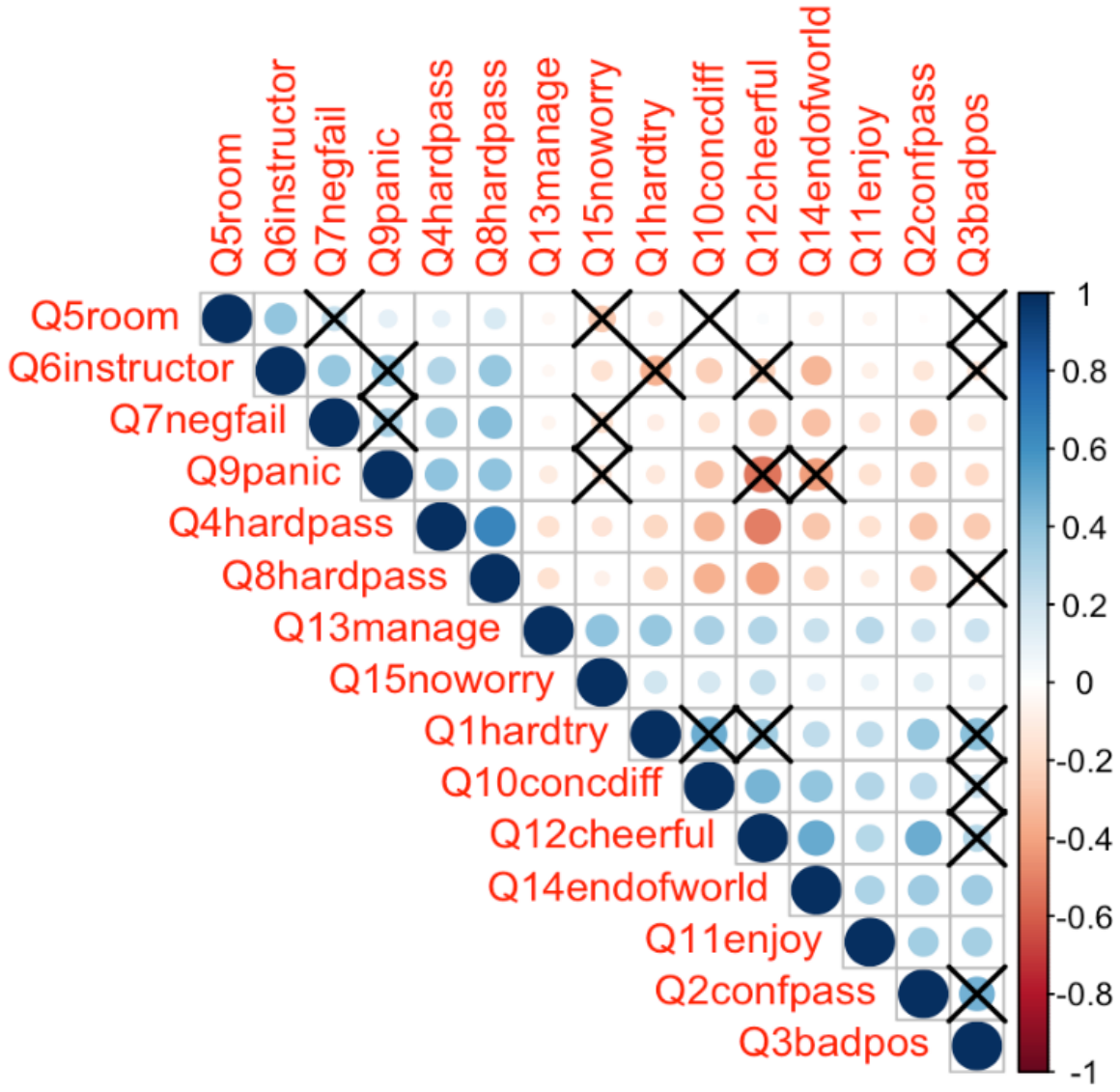


Figure H.15

Correlations for the Self-efficacy in Test Taking Instrument with the Pre-test Intervention Group, n = 59

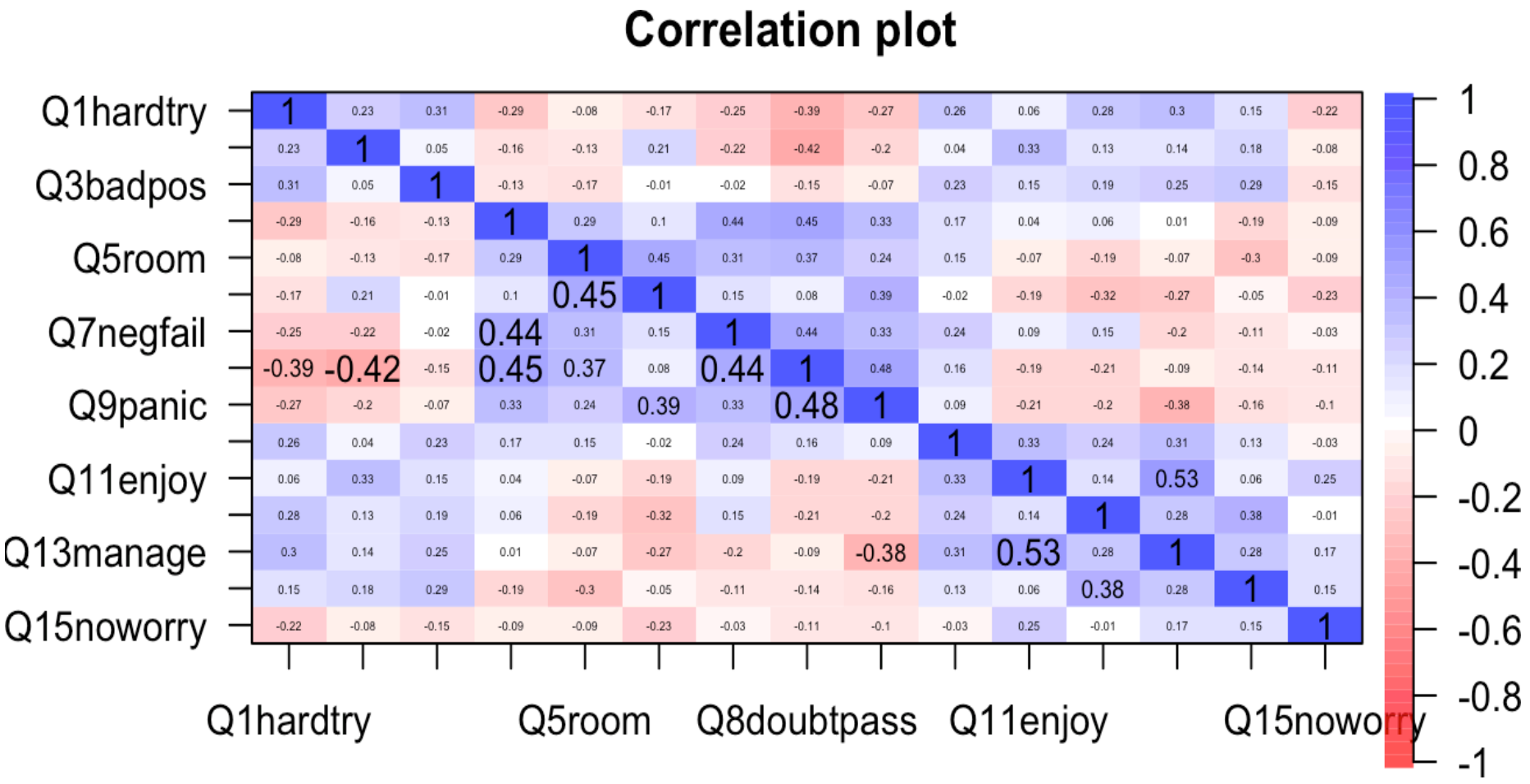


Figure H.16

Correlations for the Self-efficacy in Test Taking Instrument with the Post-test Intervention Group, n = 59

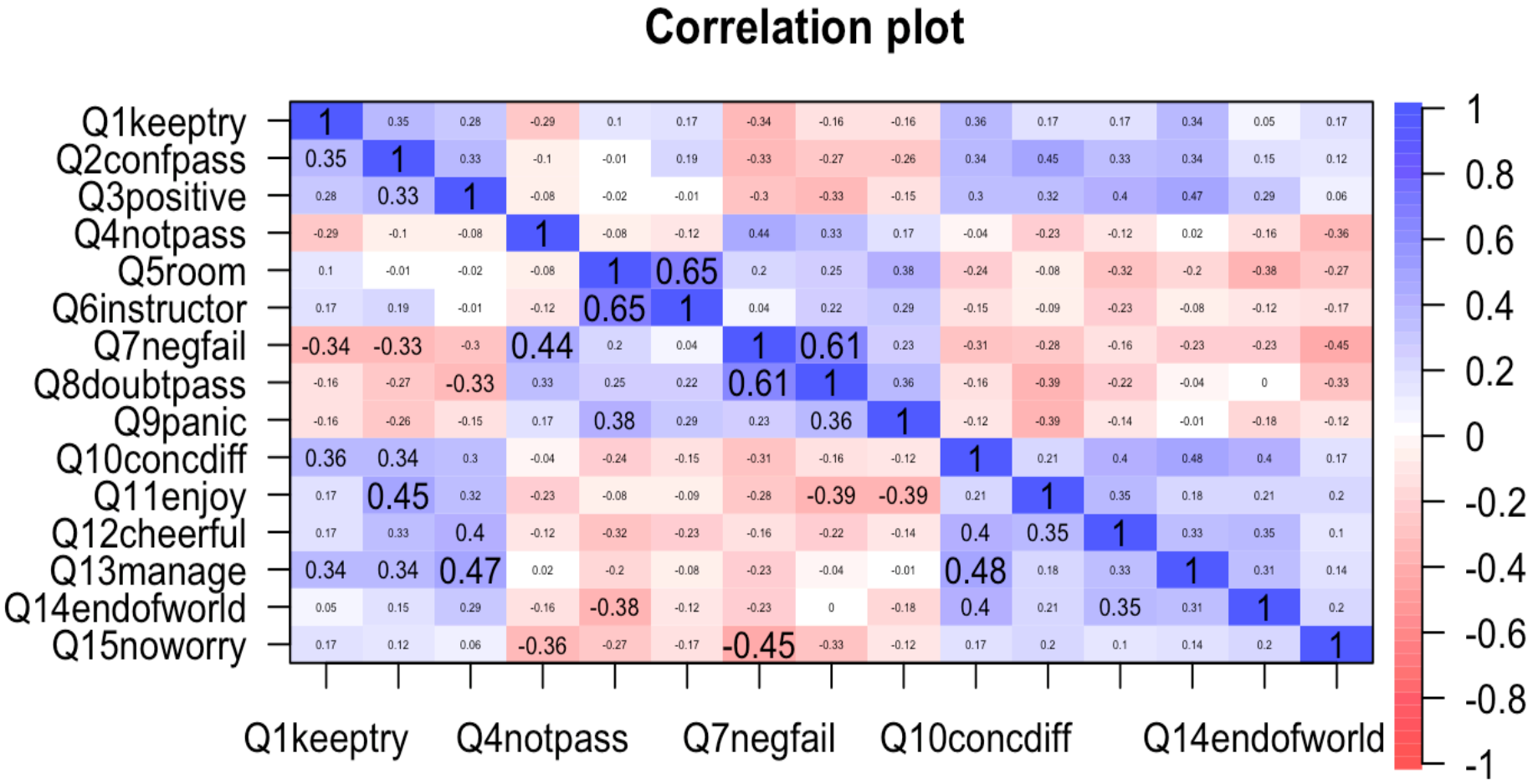


Figure H.17

Correlations for the Self-efficacy in Test Taking Instrument with the Pre-test Control Group, n = 86

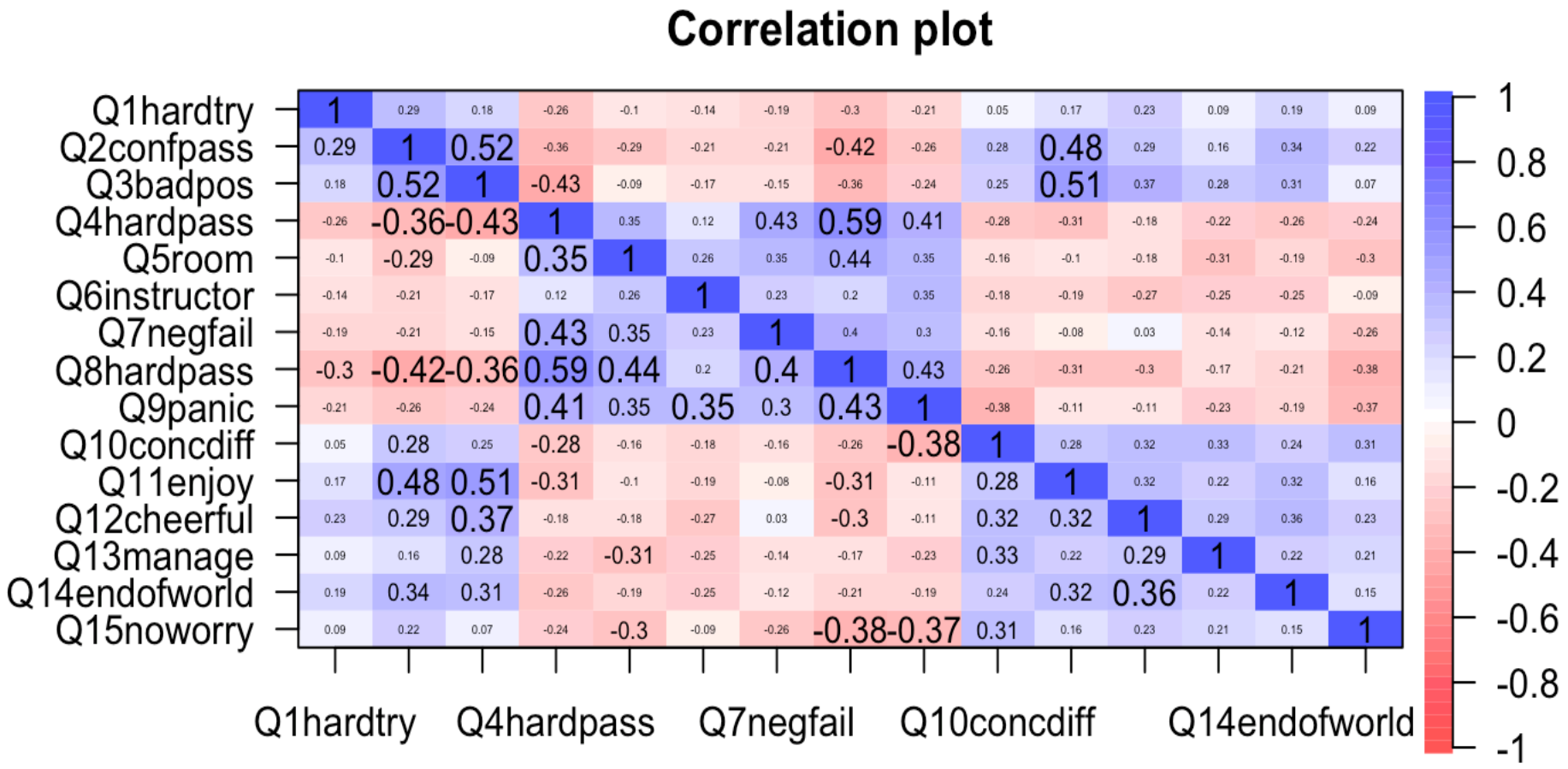
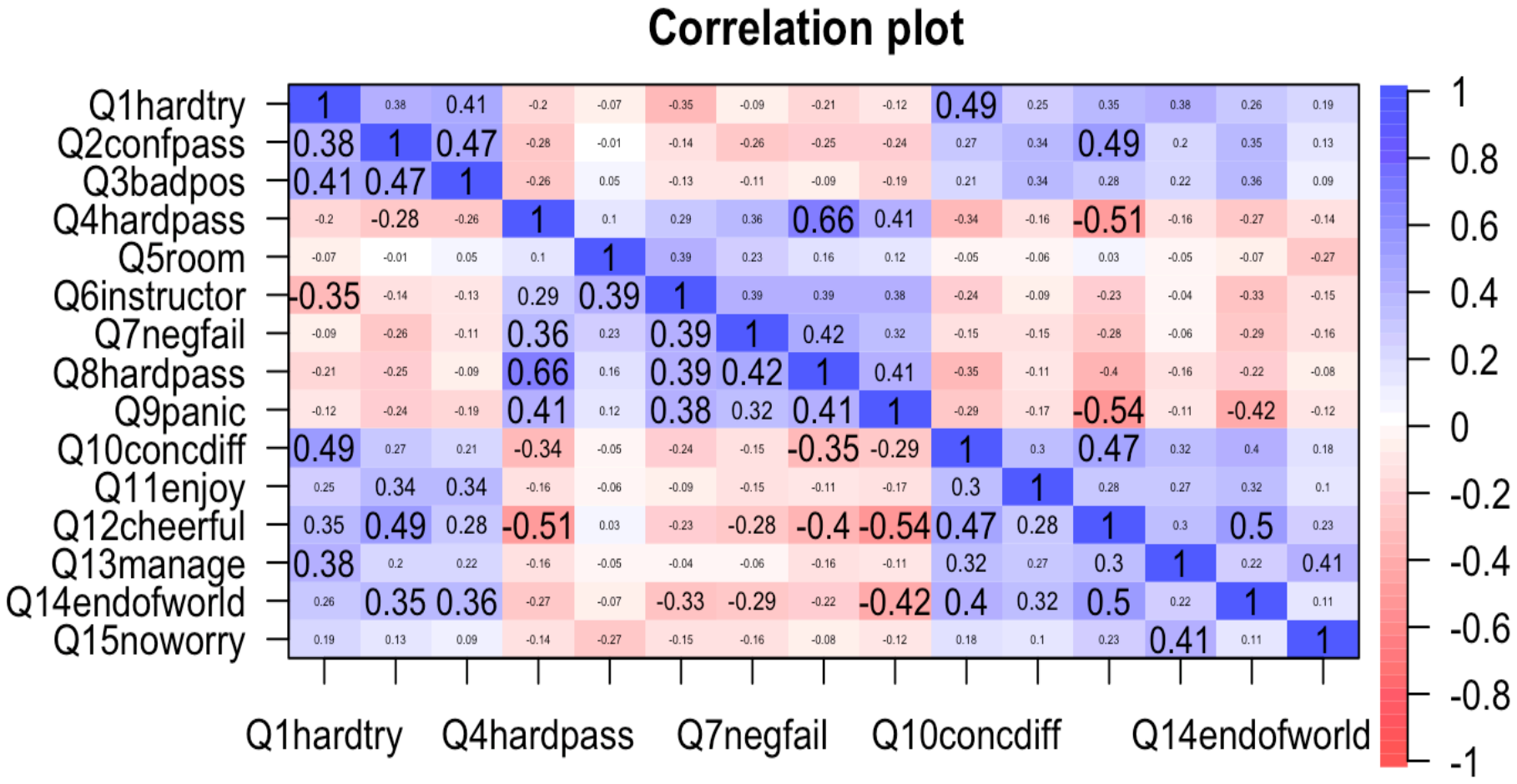


Figure H.18

Correlations for the Self-efficacy Instrument with the Post-test Control Group, n = 86



Exploratory Factor Analysis

Table H.10

Factor loadings of items onto one and two factor models for groups and timepoints for the self-efficacy data. (N = 290)

Item	Conditions							
	1	2	3	4	5	6	7	8
Q1	−0.50	−0.32 / 0.30	**0.44**	**0.60**/0.26	**0.36**	−0.24/0.26	**0.51**	**0.65**
Q2	−0.40	−0.31/0.15	**0.55**	**0.67**/0.16	**0.63**	−0.28/**0.63**	**0.56**	**0.63**
Q3	−0.26	/ **0.33**	**0.55**	**0.63**/0.14	**0.58**	−0.13/**0.73**	**0.44**	**0.66** / 0.18
Q4	**0.52**	**0.68** / 0.19	−0.34	−0.45/0.20	−0.67	**0.59**/−0.37	−0.63	**/0.71**
Q5	**0.48**	**0.48** /	−0.35	0.17/**1.05**	−0.49	**0.59**/−0.10	−0.15	/0.27
Q6	**0.34**	0.17 / −0.28	−0.19	0.21/**0.72**	−0.37	0.29/−0.23	−0.46	/**0.45**
Q7	**0.52**	**0.68**/0.15	−0.64	−0.62	−0.44	**0.58**/0.00	−0.45	/**0.52**
Q8	**0.72**	**0.76**/	−0.56	−0.49/0.13	−0.71	**0.66**/−0.36	−0.59	/**0.86**
Q9	**0.64**	**0.52**/−0.23	−0.41	−0.25/0.32	−0.55	**0.60**/−0.18	−0.60	/**0.46**
Q10	0.02	**0.42**/0.61	**0.56**	**0.50**/−0.11	**0.47**	−0.30/**0.35**	**0.59**	−0.21/−0.15
Q11	−0.28	/**0.63**	**0.55**	**0.59**	**0.51**	/**0.70**	**0.40**	**0.52**/0.13
Q12	−0.31	/**0.45**	**0.54**	**0.42**/−0.21	**0.45**	−0.13/**0.51**	**0.76**	**0.56**/−0.28
Q13	−0.40	/**0.75**	**0.49**	**0.46**	**0.40**	−0.25/**0.31**	**0.38**	**0.56**/
Q14	−0.34	−0.16 / 0.31	**0.43**	0.27/−0.31	**0.44**	−0.17/**0.45**	**0.61**	**0.51**/
Q15	−0.12	/0.12	**0.40**	0.29/−0.20	**0.43**	0.13/−**0.48**	**0.28**	**0.28**

Note. Conditions

Intervention Group (*n* = 59)
1= Pre-test x1 factor
2 = Pre-test x2 factors
3 = Post-test x1 factor
4 = Post-test x2 factors

Control Group (*n* = 85)
5 = Pre-test x1 factor
6 = Pre-test x2 factors
7 = Post-test x1 factor
8 = Post-test x2 factors

Measurement Invariance for the Self-efficacy in Test Taking Instrument Subscales: SETT (Self-efficacy in Test Taking) and LSETT (Little Self-efficacy in Test Taking)

Table H.11

Measurement Invariance for the Self-efficacy Test-Taking Subscale SETT Across Timepoint Pre-test and Post-test. (*N = 290)*

Model	Equal	X^2	*df*	*p*	*RMSEA*	*CFI*	*ΔCFI*	*TLI*	*SRMR*
Configural	Pattern	72.46	54	.02	0.05	0.95		0.94	0.05
Fully-constrained	Equal	99.30	80	.07	0.04	0.95	0.00	0.96	0.07

Note. X^2 *= chi-square value, df = Degrees of Freedom; RMSEA = root mean square error of approximation; CFI comparative fit indices; TLI = Tucker-Lewis Index; SRMR = Standardised Root Mean Square Residual* using *Bentler's* formula

Table H.12

ANOVA Between the Configural and Fully Constrained Models for the SETT subscale Across Timepoint, Pre-test and Post-test. (*N = 290)*

Model	*df*	*AIC*	*BIC*	X^2	*ΔX²*	*Δdf*	Pr (> X^2)
Configural	54	6798.6	6996.7	79.09			
Fully-constrained	80	6770.2	6872.9	102.70	25.79	26	0.47

Note. df = Degrees of Freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Pr (> X^2) is probability of being greater than *chi-square*

This non-significant result indicates that the configural and constrained (SETT, Self-efficacy in Test Taking subscale) models are the same for the timepoint groups and therefore there was no further need to test the model for metric or scalar invariance across timepoint group as the null hypothesis that the models are the same is not rejected.

Table H.13

Measurement Invariance for the Self-efficacy Test-Taking Subscale SETT Across Intervention and Control Groups. (*N = 290)*

Model	Equal	X^2	*df*	*p*	*RMSEA*	*CFI*	*⊿CFI*	*TLI*	*SRMR*
Configural	Pattern	68.66	54	0.09	0.05	0.96		0.95	0.05
Fully-constrained	Equal	119.72	80	0.003	0.04	0.90	0.06	0.91	0.09
Metric	Loadings	104.20	72	0.008	0.06	0.92	0.02	0.92	0.06
Partial Metric (Q11)	Loadings	84.37	71	0.13	0.04	0.96	0.04	0.96	0.06
Scalar				0.008	0.04	0.96	0.00	0.96	0.08

Note. ChiSq= chi-square value, df = Degrees of Freedom; RMSEA = root mean square error of approximation; CFI comparative fit indices; TLI = Tucker-Lewis Index; SRMR=Standardised Root Mean Square Residual using *Bentler's* formula.

Measurement invariance for the intervention and control groups holds from partial metric, controlling for Question 11, to scalar as the *⊿CFI* is 0.00 from partial metric to scalar.

Table H.14

ANOVA Between the Configural and Constrained Models for the SETT subscale Across Intervention and Control Groups. (*N = 290)*

Model	*df*	*AIC*	*BIC*	X^2	ΔX^2	Δdf	Pr (> X^2)
Configural	54	6778.6	6976.8	75.20			
Constrained	80	6772.8	6875.6	121.44	54.62	26	< .001
Metric	72	6775.9	6908.0	108.54	16.78	8	.03
Partial Metric	71	6758.7	6894.5	89.31		1	No value
Scalar	72	6760.9	6893.0	93.53		1	No value

Note. df = Degrees of Freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Pr (> X^2) is probability of being greater than *chi-square*

Note. Δ*CFI* taken for measurement invariance

Table H.15

Measurement Invariance for the Self-efficacy Test-Taking Subscale SETT Across Gender Groups. (N = 290)

Model	Equal	X^2	*df*	*p*	*RMSEA*	*CFI*	Δ*CFI*	*TLI*	*SRMR*
Configural	Pattern	86.21	54	.003	0.06	0.92		0.90	0.06
Fully-constrained	Equal	135.07	80	.00	0.07	0.87	0.05	0.88	0.09
Metric	Loadings	119.16	72	.00	0.07	0.89	0.02	0.89	0.07
Partial Metric	Loadings, Controlling for Q10	118.32	71	.00	0.07	0.89	0.00	0.88	0.07
Scalar		115.53	72	.00	0.07	0.89	0.00	0.89	0.08

Note. ChiSq= chi-square value, df = Degrees of Freedom; RMSEA = root mean square error of approximation; CFI = comparative fit indices; TLI = Tucker-Lewis Index; SRMR = Standardised Root Mean Square Residual using *Bentler's* formula.

Table H.16

ANOVA Between the Configural and Constrained for the SETT Subscale Across Gender Groups. (N = 290)

Model	*df*	*AIC*	*BIC*	X^2	ΔX^2	Δdf	Pr (> X^2)
Configural	54	6722.7	6920.5	91.66			
Constrained	80	6721.9	6824.5	142.90	48.93	26	< .001
Metric	72	6721.9	6853.7	126.84	16.06	8	.04
Partial Metric Q10	71	6723.0	6858.5	125.93	0.84	1	.36
Scalar	72	6719.6	6851.4	124.52	0.69	1	1

Note. df = Degrees of Freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Pr (> X^2) is probability of being greater than *chi-square*

Note. Δ*CFI* taken for measurement invariance

Table H.17

Measurement Invariance for the Little Self-efficacy in Test-Taking Subscale Across Timepoint, Pre-test and Post-test. (N = 290)

Model	Equal	X^2	*df*	*p*	*RMSEA*	*CFI*	Δ*CFI*	*TLI*	*SRMR*
Configural	Pattern	1.03	4	.90	0.00	1.00		1.05	0.007
Fully-constrained	Equal	10.28	15	.80	0.00	1.00	0.00	1.08	0.07

Note. ChiSq= chi-square value, df = Degrees of Freedom; RMSEA = root mean square error of approximation; CFI = comparative fit indices; TLI = Tucker-Lewis Index; SRMR = Standardised Root Mean Square Residual using Bentler's formula

Table H.18

ANOVA Between the Configural and Fully Constrained Models for the LSETT Subscale Across Timepoint Pre-test and Post-test. (N = 290)

Model	*df*	*AIC*	*BIC*	X^2	ΔX^2	Δdf	Pr (> X^2)
Configural	4	3042.1	3130.2	1.03			
Fully-Constrained	15	3030.2	3078.0	11.18	9.07	11	.62

Note. *df = Degrees of Freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Pr* (> X^2) is probability of being greater than *chi-square*

Note. Δ*CFI* taken for measurement invariance

The non-significant result indicates that the configural and constrained (LSETT, Little Self-efficacy in Test Taking) models are the same and therefore there was no further need to test the model for metric or scalar, or strict factorial and error invariance across timepoint groups as the null hypothesis that the models are the same is not rejected.

Table H.19

Measurement Invariance for the Little Self-efficacy Test-Taking Subscale Across Intervention and Control Groups. (N = 290)

Model	Equal	X^2	*df*	*p*	*RMSEA*	*CFI*	*ΔCFI*	*TLI*	*SRMR*
Configural	Pattern	3.10	4	.54	0.00	1.00		1.01	0.02
Fully-constrained	Equal	26.56	15	.03	0.08	0.95	0.05	0.96	0.06
Metric	Loadings	23.99	12	.02	0.09	0.94	0.01	0.94	0.06

Note. ChiSq= chi-square value, df = Degrees of Freedom; RMSEA = root mean square error of approximation; CFI comparative fit indices; TLI = Tucker-Lewis Index; SRMR = Standardised Root Mean Square Residual using *Bentler's* formula.

Measurement invariance holds across the constrained to metric models as *ΔCFI* is < 0.01 across the intervention and control groups at these stages.

Table H.20

ANOVA Between the Configural and Fully Constrained Models for the LSETT subscale Across Intervention and Control Groups. (N = 290)

Model	*df*	*AIC*	*BIC*	X^2	*ΔX²*	*Δdf*	Pr (> X^2)
Configural	4	3031.1	3119.2	3.24			
Constrained	15	3034.6	3082.3	28.71	23.26	11	.02
Metric	12	3040.3	3099.0	28.42	0.29	3	.93

Note. df = Degrees of Freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Pr (> X^2) is probability of being greater than *chi-square*

Note. ΔCFI taken for measurement invariance

Table H.21

ANOVA Between the Constrained and Metric Models for the LSETT subscale Across Intervention and Control Groups. (N = 290)

Model	*df*	*AIC*	*BIC*	X^2	ΔX^2	Δdf	Pr (> X^2)
Metric	12	3040.3	3099.0	28.42			
Constrained	15	3034.6	3082.3	28.72	0.44	3	.93

Note. *df = Degrees of Freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Pr* (> X^2) is probability of being greater than *chi-square*

Note. Δ*CFI* taken for measurement invariance

The significant value between constrained and configural indicates that the null hypothesis that both models are the same is rejected and the model is tested for metric and scalar measurement invariance. Measurement invariance holds from the constrained through to metric models.

Table H.22

Measurement Invariance for the Little Self-efficacy Test-Taking Subscale Across Gender Groups. (*N = 290)*

Model	Equal	X^2	*df*	*p*	*RMSEA*	*CFI*	Δ*CFI*	*TLI*	*SRMR*
Configural	Pattern	3.37	4	.50	0.00	1.00		1.00	0.02
Fully-constrained	Equal	17.28	15	.30	0.03	0.99	0.01	0.99	0.09

Note. ChiSq= chi-square value, df = Degrees of Freedom; RMSEA = root mean square error of approximation; CFI comparative fit indices; TLI = Tucker-Lewis Index; SRMR = Standardised Root Mean Square Residual using *Bentler's* formula.

Measurement invariance holds across the configural to constrained models at these stages for gender groups as Δ*CFI* does not exceed < 0.01.

Table H.23

ANOVA Between the Configural and Fully Constrained Models for the LSETT subscale Across Gender Groups. (N = 290)

Model	*df*	*AIC*	*BIC*	X^2	ΔX^2	Δdf	Pr (> X^2)
Configural	4	3018.8	3106.7	3.25			
Constrained	15	3011.5	3059.1	17.95	13.79	11	.25

Note. df = Degrees of Freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Pr (> X^2) is probability of being greater than *chi-square*

Note. Δ*CFI* taken for measurement invariance

ANOVA tables have been provided for all measurement invariance calculations across stages for the Self-efficacy in Test Taking subscale and for Non-Self-efficacy in Test Taking subscale.

Research Question 2a: Does the intervention increase feelings of self-efficacy in test taking in 10-to-11-year-olds?

Assumptions for the Self-efficacy in Test Taking Subscale

Table H.24

Levene's Test for Homogeneity of Variance - Self-efficacy in Test Taking

Timepoint	*F*	*df*1	*df*2	*p*
Pre	1.45	1	143	.23
Post	0.05	1	143	.83

Mixed 2 x 2 Factorial ANOVA for the Self-efficacy in Test Taking Instrument

Table H.25

Results for Within Subjects Effects for Timepoint for the Self-efficacy in Test Taking Instrument (N = 290)

Group	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2	η^2_p
Timepoint	49.11	1	49.11	5.25	.02	0.007	0.04
Timepoint*Group	1.15	1	1.15	0.12	.73	< .001	< .001
Residuals	1336.83	143	9.35				

Note: Type III Sum of squares

Table H.26

Results for Between Subjects Effects for Group for the Self-efficacy in Test Taking Instrument (N = 290)

Group	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2	η^2_p
Group	9.69	1	9.69	0.24	.62	0.001	0.002
Residuals	5722.73	143	40.02				

Note: Type III Sum of squares

Table H.27

Descriptive Statistics for the Groups for the Self-Efficacy in Test Taking Instrument (N =290)

Timepoint	Group	*Mean*	*SD*	*N*
Pre-test	Intervention	23.81	4.32	59
	Control	24.31	5.06	86
Post-test	Intervention	24.78	4.99	59
	Control	24.78	5.26	86

Latent Means Analyses

Table H.28

Latent Means Fit Statistics of Self-efficacy in Test Taking Subscale Across the Timepoints Pre and Post for Intervention (n = 118)

Model	*X²*	*df*	*CFI*	*TLI*	*RMSEA*	*SRMR*
Equal means	79.06	81	1.00	1.00	0.00	0.10
Unequal means	77.44	80	1.00	1.02	0.00	0.10

Note. X² = chi-square value, df = Degrees of Freedom; RMSEA = root mean square error of approximation; CFI comparative fit indices; TLI = Tucker-Lewis Index; SRMR = Standardised Root Mean Square Residual using *Bentler's* formula.

Table H.29

Latent Means'ANOVA for Self-efficacy in Test Taking Across the Timepoints Pre and Post Intervention Groups (n=118)

Model	*df*	*AIC*	*BIC*	*X²*	*ΔX²*	*Δdf*	Pr (> *X²*)
Equal Means	81	2751.5	2826.3	82.18	1.96	1	.16
Unequal Means	80	2752.1	2829.7	80.83			

Note. df = Degrees of Freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Pr (> *X²*) is probability of being greater than *chi-square*

Table H.30

Latent Means' Fit Statistics of Self-efficacy in Test Taking Subscale Across the Timepoints Pre and Post Control Groups (n =172).

Model	X^2	*df*	*CFI*	*TLI*	*RMSEA*	*SRMR*
Equal means	92.90	81	0.95	0.96	0.04	0.08
Unequal means	91.33	80	0.95	0.96	0.04	0.08

Note. ChiSq = chi-square value, df = Degrees of Freedom; CFI comparative fit indices; TLI = Tucker-Lewis Index; RMSEA = root mean square error of approximation; SRMR = Standardised Root Mean Square Residual using *Bentler's* formula.

Table H.31

Latent Means' ANOVA for the Self-efficacy in Test Taking Subscale Across the Timepoints Pre and Post for Control Groups (n=172)

Model	*df*	*AIC*	*BIC*	X^2	ΔX^2	Δdf	Pr (> X^2)
Equal Means	81	4027.1	4112.1	93.69	1.65	1	.20
Unequal Means	80	4027.7	4115.8	92.29			

Note. df = Degrees of Freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Pr (> X^2) is probability of being greater than *chi-square*

Mixed 2 x 2 Factorial ANOVA With Little Self-efficacy (LSETT) Scale

Assumptions

Table H.32

Levene's Test for Homogeneity of Variance - Little Self-efficacy in Test Taking

Timepoint	*F*	*df*1	*df*2	*p*
Pre	0.83	1	143	.36
Post	0.25	1	143	.62

Table H.33

Results for Within Subjects Effects for Timepoint for the Little Self-efficacy in Test Taking Instrument (N = 290)

Group	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2	η^2_p
Timepoint	54.44	1	54.44	18.19	< .001	0.02	0.11
Timepoint* Group	13.56	1	13.56	4.53	.04	0.00	0.03
Residuals	1159.78	143	9.998				

Note: Type III Sum of squares

Table H.34

Results for Between Subjects Effects for Group for the Little Self-efficacy in Test Taking Instrument (N = 290)

Group	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2	η^2_p
Group	16.01	1	16.01	1.13	.29	0.006	0.008
Residuals	2016.07	143	14.10				

Note: Type III Sum of squares

Table H.35

Descriptive Statistics for the Groups for the Little Self-Efficacy in Test Taking Instrument (N = 290)

Timepoint	Group	*Mean*	*SD*	*N*
Pre-test	Intervention	9.41	4.32	59
	Control	8.49	5.01	86
Post-test	Intervention	8.09	4.99	59
	Control	8.05	5.17	86

Table H.36

Post Hoc Test for Little Self-efficacy Subscale (N = 290)

Group/ Timepoint	Group/ Timepoint	Mean Difference	*SE*	*t*	*p Tukey*
intervention, pre	control, pre	0.92	0.49	1.86	0.25
	intervention, post	**1.32**	**0.32**	**4.15**	**< .001**
	control, post	1.36	0.49	2.75	0.03
control, pre	intervention, post	0.40	0.49	0.82	0.85
	control, post	0.44	0.26	1.68	0.34
intervention, post	control, post	0.04	0.49	0.07	1.00

Note. p value adjusted for comparing a family of 6. **Bold** font indicates the significant value taken.

ANCOVA

Table H.37

Little Self-efficacy in Test Taking Levene's Test – ANCOVA (N = 145)

Timepoint	*F*	*df*1	*df*2	*p*
Post	3.35	1	143	.07

Table H.38

Little Self-efficacy in Test Taking Subscale ANCOVA (N = 145)

Cases	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2	η^2_p
Pre centred	467.09	1	467.09	105.74	< .001	0.43	0.43
Group	8.63	1	8.63	1.95	.16	0.008	0.014
Residuals	627.29	142	4.418				

Note: Type I Sum of squares.

Table H.39

Little Self-efficacy in Test Taking Subscale Marginal Means

Group	*Marginal Mean*	95% Confidence		*SE*	*t*	*df*	*p*
		Lower	Upper				
Intervention	7.76	7.22	8.31	0.28	28.19	142	< .001
Control	8.23	7.82	8.72	0.23	36.31	142	< .001

Note: Confidence refers to Confidence Interval for Mean Difference.

Latent Means Analyses

Table H.40

Latent Means' Fit Statistics for Little Self-efficacy in Test Taking Subscale Across the Timepoints Pre and Post for the Intervention Group (n =118)

Model	X^2	*df*	*CFI*	*TLI*	*RMSEA*	*SRMR*
Equal means	19.97	16	0.95	0.96	0.06	0.14
Unequal means	14.70	15	1.00	1.00	0.00	0.10

Note. X^2 *= chi-square value, df = Degrees of Freedom; RMSEA = root mean square error of approximation; CFI comparative fit indices; TLI = Tucker-Lewis Index; SRMR = Standardised Root Mean Square Residual* using *Bentler's* formula.

Table H.41

Latent Means' ANOVA for the Little Self-efficacy in Test Taking Subscale Across the Timepoints Pre and Post for the Intervention Group (n =118)

Model	*df*	*AIC*	*BIC*	X^2	ΔX^2	Δdf	Pr (> X^2)
Equal Means	16	1253.0	1286.2	21.46	4.95	1	< .001
Unequal Means	15	1250.0	1286.0	16.51			

Note. df = Degrees of Freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Pr (> X^2) is probability of being greater than *chi-square*

Table H.42

Latent Means' Fit Statistics of Little Self-efficacy in Test Taking' Subscale Across the Timepoints Pre and Post Control Group (N =172)

Model	X^2	*df*	*CFI*	*TLI*	*RMSEA*	*SRMR*
Equal means	13.66	16	1.00	1.01	0.00	0.08
Unequal means	12.23	15	1.00	1.01	0.00	0.08

Note. X^2 *= chi-square value, df = Degrees of Freedom; CFI comparative fit indices; TLI = Tucker-Lewis Index; RMSEA = root mean square error of approximation; SRMR = Standardised Root Mean Square Residual* using *Bentler's* formula.

Table H.43

Latent Means' ANOVA for the Little Self-efficacy in Test Taking Subscale Across the Timepoints Pre and Post for the Control Group (N =172)

Model	*df*	*AIC*	*BIC*	X^2	ΔX^2	Δdf	Pr (> X^2)
Equal Means	16	1778.1	1815.9	14.37	1.45	1	.23
Unequal Means	15	1778.7	1815.6	12.91			

Note. df = Degrees of Freedom; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; Pr (> X^2) is probability of being greater than *chi-square*

Research Question 3: Do feelings of worry predict how well student self-efficacy changes in the intervention group?

Mixed 2 x 2 Factorial ANOVAs For High Test Anxiety Students

Assumptions

Table H.44

Levene's Test for Homogeneity of Variance for High Test Anxiety Students Measured with the LSETT (Little Self-efficacy in Test Taking) Subscale

Timepoint	*F*	*df*1	*df*2	*p*
Pre	3.256	1	29	.08
Post	1.697	1	29	.20

Table H.45

Results for Within Subjects Effects for Timepoint for the Little Self-efficacy in Test Taking Subscale with Students Who Are High in Test Anxiety (N = 31)

Group	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2	η^2_p
Timepoint	4.759	1	4.759	2.068	.16	0.009	0.07
Timepoint* Group	12.824	1	12.824	5.573	.03	0.024	0.16
Residuals	66.725	29	2.301				

Note: Type III Sum of squares

Table H.46

Results for Between Subjects Effects for Group for the Little Self-efficacy in Test Taking Subscale with Students Who Are High in Test Anxiety (N = 31)

Group	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2	η^2_p
Group	29.431	1	29.431	1.977	.17	0.05	0.06
Residuals	431.666	29	14.885				

Note: Type III Sum of squares

Table H.47

Descriptive Statistics for the Groups for the Little Self-Efficacy in Test Taking Instrument with Students Who Are High in Test Anxiety (N = 31)

Timepoint	Group	*Mean*	*SD*	*N*
Level 1 (Pre-test)	Intervention	9.94	3.56	17
	Control	7.64	2.53	14
Level 2 (Post-test)	Intervention	8.47	3.04	17
	Control	8.00	2.22	14

Post Hoc Test

Table H.48

Post hoc Test for the Interaction Between Group and Timepoint – Little Self-efficacy Subscale (N =31)

Group/ Timepoint	Group/ Timepoint	Mean Difference	*SE*	*t*	*p* Tukey
intervention, pre	control, pre	2.30	1.06	2.17	.15
	intervention, post	**1.47**	**0.52**	**2.83**	**.04**
	control, post	1.94	1.06	0.27	.30
control, pre	intervention, post	−0.83	1.06	−0.78	.86
	control, post	−0.36	0.57	−0.62	.92
intervention, post	control, post	0.47	1.06	0.45	1.00

Note. **Bold** font indicates the significant value taken.

Mixed 2 x 2 ANOVA for Self-efficacy in Test Taking Subscale

Assumptions

Table H.49

Levene's Test for Homogeneity of Variance for High Test Anxiety Students Measured with the SETT (Self-efficacy in Test Taking) Subscale

Timepoint	*F*	*df*1	*df*2	*p*
Pre	0.11	1	29	.74
Post	0.95	1	29	.34

Table H.50

Results for Within Subjects Effects for Timepoint for the Self-efficacy in Test Taking Subscale with Students Who Are High in Test Anxiety (N = 31)

Group	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2	η^2_p
Timepoint	1.25	1	1.25	0.15	.70	< .01	0.005
Timepoint* Group	7.83	1	7.83	0.94	.34	.005	0.03
Residuals	240.71	29	8.30				

Note: Type III Sum of squares

Table H.51

Results for Between Subjects Effects for Group for the Self-efficacy in Test Taking Subscale with Students Who Are High in Test Anxiety (N = 31)

Group	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2	η^2_p
Group	268.91	1	268.91	7.36	.01	0.17	0.20
Residuals	1059.19	29	36.52				

Note: Type III Sum of squares

Table H.52

Descriptive Statistics for the Groups for the Self-Efficacy in Test Taking Instrument with Students Who Are High in Test Anxiety (N = 31)

Timepoint	Group	*Mean*	*SD*	*N*
Level 1 (Pre-test)	Intervention	21.53	4.74	17
	Control	26.43	4.99	14
Level 2 (Post-test)	Intervention	22.56	5.19	17
	Control	26.00	3.78	14

Mixed 2 x 2 ANOVA with Adapted Thoughts

Assumptions

Table H.53

Levene's Test for Homogeneity of Variance for High Test Anxiety Students Measured with the Adapted Thoughts' Subscale

Timepoint	*F*	*df*1	*df*2	*p*
Pre	0.11	1	29	.74
Post	0.73	1	29	.40

Table H.54

Results for Within Subjects Effects for Timepoint for the Adapted Thoughts' Subscale with Students Who Are High in Test Anxiety (N = 31)

Group	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2	η^2_p
Timepoint	88.54	1	88.54	7.75	.009	0.05	0.21
Timepoint* Group	21.63	1	21.64	1.89	.18	0.01	0.06
Residuals	331.24	29	11.42				

Note: Type III Sum of squares

Table H.55

Results for Between Subjects Effects for Group for the Adapted Thoughts' Subscale with Students Who Are High in Test Anxiety (N = 31)

Group	Sum of Squares	*df*	Mean Square	*F*	*p*	η^2	η^2_p
Group	83.05	1	83.05	2.00	.17	0.05	0.06
Residuals	1205.37	29	41.57				

Note: Type III Sum of squares

Table H.56

Descriptive Statistics for the Groups for Adapted Thoughts' Subscale with Students Who Are High in Test Anxiety (N = 31)

Timepoint	Group	*Mean*	*SD*	*N*
Level 1 (Pre-test)	Intervention	22.65	4.23	17
	Control	23.79	4.02	14
Level 2 (Post-test)	Intervention	19.06	6.37	17
	Control	22.57	5.46	14

Appendix I

R Studio© Code

Samples of Measurement Invariance for Self-efficacy and Children's Test Anxiety Scale

```
masterynonsefinequal<-cfa(SEsubscalenonefficacywithoutskew,
data=AdditionalSE_data_complete[,1:19], group = "pre_post", group.equal=c("loadings",
"intercepts", "residuals", "lv.variances", "lv.covariances"), estimator="MLM")#NSETT
models

summary (masterynonseequal, fit.measures=TRUE, standardized=TRUE)

anova(masterynonsefin, masterynonsefinequal)

mastery.mif.equal<-cfa(MIS1F,data=FFinal_Pre_post_int_non_int[,2:35],
group="Timepoint", group.equal=c("intercepts", "residuals", "lv.variances",
"lv.covariances"), estimator="MLM")

summary(mastery.mif.equal, standardized=TRUE, fit.measures=TRUE)# metric model

mastery.mif.equal.plot<-semPaths(mastery.mif.equal, what="col", whatLabels = "std",
style="lisrel", layout = "tree")
```

Mimic Models

```
mastery.mimic.model.post.intervention.SETT<-
'mastery=~Q1+Q5+Q9+Q11+Q16+Q19+Q27+Q28
mastery~SETT'

mastery.mimic.model.post.intervention.SETT.fit<-
sem(mastery.mimic.model.post.intervention.SETT,
data=ATINterventionPOSTsett_n_nsett,estimator="MLM")

summary(mastery.mimic.model.post.intervention.SETT.fit, standardized=T, fit.measures=T)

modcont5<-semPaths(mastery.mimic.model.post.intervention.SETT.fit, what="mod",
whatLabels = "Parameters", style="OpenMx",edge.width=0.5,.edge.color="black", sizeMan
= 8, sizeLat = 20)
```

Latent Means after CFA

```
mastery.model.equal.means.final<-cfa(MIS1F, data=FFinal_Pre_post_int_non_int[,2:35],
group="Timepoint", group.equal=c("loadings", "intercepts", "residuals", "lv.variances",
"means"), estimator="MLM")

summary(mastery.model.equal.means.final, fit.measures=T, standardized=T)# model fit the
data well -for the constrained model -both group means are set to zero –
```

```
mastery.model.unequal.means.final<-cfa(MIS1F, data=FFinal_Pre_post_int_non_int[,2:35],
group="Timepoint", group.equal=c("loadings", "intercepts", "residuals", "lv.variances"),
estimator="MLM")

summary(mastery.model.unequal.means.final, standardized=T, fit.measures=T)
# test for means difference is statistically significant –

anova(mastery.model.equal.means.final, mastery.model.unequal.means.final)# can I do this
for pre post intervention and control -
```

Playing with Plots

```
M<-cor(AIPCSSwithlabels[,6:35])#creating an object to work with! Pre-test intervention
M
corrplot(M, method="circle")# ok so works but need sig values
corrplot(M, method="number")
corrplot(M, method="color")
corrplot(M, order="AOE")
corrplot(M, order="hclust", addrect=2)
corrplot(M, p.mat=res_1$p,  sig.level=.5, type="upper", order="hclust", tl.cex = 0.5)


A<-cor(FINSE_NONINT_PRE_lables[,4:18])
corrplot (A, p.mat=res_1$p, sig.level=.5, type="upper", order="hclust")

B<-cor(FINSE_NON_INT_POST_lables[,4:18])
corrplot (B, p.mat=res_1$p, sig.level=.5, type="upper", order="hclust")
# here goes with the SE data  -correlation matrices.

P<-cor(ANIPOFSS_with_labels[,5:34])# so this gave pretty fab plots but left sig values off
det(P)
Q<-cor(SE_PRE_INT_labels[,4:18])
corrplot(Q, p.mat+res_1$p, sig.level=.5, type="upper", order="hclust")

det(R) # (determinant values)

T<-cor(SE_NON_INT_POST_label[,4:18])

cg<-corrplot(M, type="lower", order="hclust", col=c("black", "white"), bg="lightblue")
corrplot(cg, p.mat=res_1$p, sig.level=.5)# gives me sig values


cor.plot(ANIPOFSS[,5:34],numbers=TRUE,colors=TRUE,n=51,zlim=c(-1,1),
 show.legend=TRUE,keep.par=TRUE,select=NULL, pval=NULL,
  cuts=c(.001,.01),scale=FALSE,upper=TRUE,diag=TRUE, labels=NULL,
symmetric=TRUE,stars=FALSE,adjust="holm",xaxis=3,xlas=1,ylas=2,alpha=.75,
min.length=NULL, MAR = TRUE, cex=0.5,cex.axis=0.3)# this code works for a picture of
correlation with readable numerics.
```