UNCOVERING THE ROOT CAUSES OF ETHNIC DIFFERENCES IN ABILITY TESTING

Differential Test Functioning, Test Familiarity and Trait Optimism as Explanations of Ethnic Group Differences

DANIEL PRICE HINTON

Doctor of Philosophy

ASTON UNIVERSITY

October 2014

ASTON UNIVERSITY

UNCOVERING THE ROOT CAUSES OF ETHNIC DIFFERENCES IN ABILITY TESTING

Differential Test Functioning, Test Familiarity and Trait Optimism as Explanations of Ethnic Group Differences

DANIEL PRICE HINTON

Doctor of Philosophy

2014

THESIS SUMMARY

The present research represents a coherent approach to understanding the root causes of ethnic group differences in ability test performance. Two studies were conducted, each of which was designed to address a key knowledge gap in the ethnic bias literature.

In Study 1, both the LR Method of Differential Item Functioning (DIF) detection and Mixture Latent Variable Modelling were used to investigate the degree to which Differential Test Functioning (DTF) could explain ethnic group test performance differences in a large, previously unpublished dataset. Though mean test score differences were observed between a number of ethnic groups, neither technique was able to identify ethnic DTF. This calls into question the practical application of DTF to understanding these group differences.

Study 2 investigated whether a number of non-cognitive factors might explain ethnic group test performance differences on a variety of ability tests. Two factors – test familiarity and trait optimism – were able to explain a large proportion of ethnic group test score differences. Furthermore, test familiarity was found to mediate the relationship between socio-economic factors – particularly participant educational level and familial social status – and test performance, suggesting that test familiarity develops over time through the mechanism of exposure to ability testing in other contexts.

These findings represent a substantial contribution to the field's understanding of two key issues surrounding ethnic test performance differences. The author calls for a new line of research into these performance facilitating and debilitating factors, before recommendations are offered for practitioners to ensure fairer deployment of ability testing in high-stakes selection processes.

Keywords: Race, bias, Differential Item Functioning, mixture modelling, test familiarity

To Grandad

ACKNOWLEDGEMENTS

I would like to thank a number of people without whose guidance and support completion of this thesis would have been impossible. Firstly, thanks go to Professor Stephen Woods and Dr Yves Guillaume of Aston Business School for their constant help and support as primary and secondary supervisors for this project over the past four years.

Additionally, thanks go to Matt Stewart and the team at Pearson Talent Lens, and Roy Childs and Team Focus for agreeing to the use of their tools, for granting access to their test data archives, and for help in the use of their on-line assessment platforms.

I would like to thank a number of people who helped in the data collection process. In particular, thanks go to Virinder Hayre and Dr Andrew Clements for granting me access to participants in their respective organisations. I would also like to acknowledge the hard work of Amy Parsons, Angie Ingman, Darren Schrouder, Kelly Drane, Chris Walkley, Michael Towl, Jenny Surtees, Aaron Percival, and Harvinder Bharj for helping me to recruit participants for Study 2. Special thanks go to Kevin Teoh for his commitment to my project.

I would also like to acknowledge my two viva examiners, Dr. Anna Brown (External Examiner) and Dr. Nicholas Theodorakopoulos (Internal Examiner). In particular, the former's mastery of advanced statistical methods not only challenged my own understanding of my analyses, it helped to clarify the approach I had taken to addressing my research questions. This resulted in a much more finely-honed finished product, one which, overall, makes a more compelling argument for my position.

Finally, I would like to thank Alison Chesters for the often-saintly degree of patience that she has shown me these past years. You are a continuing source of support and inspiration for me, and I might not have had the motivation to see this through had you not been there.

LIST OF CONTENTS

Thesis Summary 2
Acknowledgements 4
List of Contents 5
List of Tables and Figures 9
Chapter 1: Introduction 13
1.1 The Present Research 14
1.1.1 Study 1 15
1.1.2 Study 2 16
Chapter 2: Literature Review 19
2.1 The Structure and Measurement of Cognitive Ability 19
2.2 The Relationship of Intelligence to Job Performance 22
2.3 Ethnic Group Differences in Ability Test Performance 23
2.4 Competing Conceptualisations of Group Differences: Test Impact and Test Bias 25
2.4.1 Adverse Impact 26
2.4.2 Cleary's Rule: Differential Prediction 28
2.4.3 Thorndike's Rule 29
2.4.4 Measurement Invariance and Item Response Theory 31
2.4.4.1 Differential Item Functioning and Differential Test Functioning 35
2.4.4.2 Mixture Latent Variable Modelling 39
2.4.4.3 The Tension between Measurement Invariance and Prediction Invariance 43
2.5 Explanations for Patterns of Group Difference 44
2.5.1 The Relationship between *g* and Test Performance in Classical Test Theory 47
2.5.2 Test Familiarity as a Performance Facilitating Factor 53
2.5.3 A New Model of Ethnic Test Performance Differences 57
2.6 Further Issues in Ethnic Bias Research 60
2.6.1 The Spearman-Jensen Effect 60
2.6.2 Range Restriction 63
2.6.3 Representing Ethnicity 64
Chapter 3: Study 1 66
3.1 Study Overview 66
3.1.1 Hypotheses 67
3.2 Raven's SPM Data Archive 68
3.2.1 Measures 68

Raven's Standard Progressive Matrices 68
Demographic Variables 71
3.2.2 Sample 71
3.2.2.1 Conceptualising Ethnicity 72
3.2.3 Results 74
3.2.3.1 Ethnic Group Test Score Differences 74
3.2.3.2 Traditional Ethnic DTF Identification using the LR Method 78
SPM Iteration 1 81
SPM White-Asian DIF Analysis – Iteration 1 84
SPM White-Middle Eastern DIF Analysis – Iteration 1 86
3.2.3.3 DTF Analysis using MLVM 88
3.2.3.4 MLVM on Unifactorial Item Subsets within the SPM 104
3.3 Attempted Replication using Raven's APM Data Archive 108
3.3.1 Method 109
3.3.1.1 Measures 109
Raven's Advanced Progressive Matrices 109
Demographic Variables 109
3.3.1.2 Sample 110
3.3.1.2.1 Reclassifying Ethnicity 110
3.3.2 Results 111
3.3.2.1 Ethnic Group Test Score Differences 111
3.3.2.2 Traditional Ethnic DTF Identification using the LR Method 113
APM Iteration 1 114
APM White-Asian DIF Analysis – Iteration 1 116
3.3.2.3 DTF Analysis using MLVM 118
3.3.2.4 MLVM on Unifactorial Item Subsets within the APM 126
3.4 Discussion 129
3.4.1 Ethnic Differences in Test Performance 129
3.4.2 Ethnic DIF and DTF in the Raven's Measures 132
Chapter 4: Study 2 141
4.1 Study Overview 141
4.1.1 Hypotheses 142
4.2 Method 148
4.2.1 Measures 148
PfS Verbal 149
PfS Numerical 149

PfS Abstract 150
Memory and Attention Test (MAT) 152
Raven's Standard Progressive Matrices (SPM) 154
Trait Personality Inventory 154
Participant and Familial Social Status 157
Other Demographic Variables 160
4.2.1.1 Designing the Test Familiarity Scale 160
4.2.2 Sample 166
4.2.2.1 Sampling Strategy 166
4.2.2.2 Conceptualising Ethnicity 167
4.2.2.3 Conceptualising Employment Status 168
4.2.2.4 Participants 169
4.2.3 Procedure 171
4.3 Results 173
4.3.1 Ethnic Mean Test Score Differences 174
4.3.2 The Variance in Test Scores Explainable through Ethnicity 176
PfS Verbal 177
PfS Numerical 177
PfS Abstract 177
Raven's SPM 178
MAT 178
4.3.3 The Nature of Performance Facilitating and Debilitating Factors 180
4.3.3.1 Socio-economic Variables and their Influence on Test Performance 180
4.3.3.2 Test Familiarity as a Performance Facilitating Factor 182
4.3.3.3 Explaining Differences across Tests in the Facilitative Effect of Test Familiarity 187
4.3.3.4 Personality Traits as Performance Facilitating and Debilitating Factors 191
4.3.5 Ethnic Differences on Key Variables 194
4.3.6 Test Familiarity Differences as an Explanation for Group Differences 198
4.3.7 Optimism as an Explanation of Ethnic Group Test Performance Differences 200
4.3.8 The Combined Effect of Performance Facilitating and Debilitating factors 202
Chapter 5: General Discussion 205
5.1 Chapter Overview 205
5.2 Summary of Key Findings 206
5.2.1 Ethnic Group Test Score Differences 206
5.2.2 Explanations for these Differences 209

5.2.2.1 Differential Test Functioning 209
5.2.2.2 Test Familiarity as a Performance Facilitating Factor 211
5.2.2.3 Personality Traits and their Influence on Performance 215
5.3 Academic Contribution 220
5.3.1 A Model of Ethnic Group Difference in Test Performance 222
5.4 Practical Impact and Recommendations for Practitioners 224
5.5 Limitations to the Present Research 226
5.6 Directions for Future Research 229
5.7 Conclusion 232
List of References 234

LIST OF TABLES AND FIGURES

Figure 1. Carroll's (1993) three-level hierarchy of intelligence (adapted from Robertson et al, 2002) 20
Figure 2. The practical impact of a 1 S.D. difference in mean test score between groups at selection. 27
Figure 3. Graph showing different regression lines for higher and lower scoring groups on the same measure of cognitive ability. 28
Figure 4. Graph showing disparity in size of mean group difference in job performance and test score for two separate social groups. 30
Figure 5. Example ICC for an item according to the 3PL model. 34
Figure 6. ICCs for an item that shows DIF. 36
Figure 7. Example ICCs that might be observed for an item that shows Non-uniform DIF. .. 38
Figure 8. Mixture latent variable model estimated by Mplus for a test consisting of 8 items (from Muthén & Muthén, 2010) 40
Figure 9. A model for understanding the personality-intelligence interface (from Chamorro-Premuzic & Furnham, 2004). 48
Figure 10. Yerkes-Dodson curve showing the relationship between stress level and task performance. 49
Figure 11. Proposed model showing how facets of *g*, facilitating and debilitating factors interact to bring about ethnic group test performance differences. 58
Figure 12. Example item from the Raven's SPM (used with kind permission from Pearson Education Ltd.) 69
Table 1. Mean raw score, percentiles and Cohen's *d* by ethnicity classified as White/Non-white. 75
Table 2. Mean raw score, percentiles and Cohen's *d* by broad ethnic group classification. . 75
Table 3. Mean raw score, percentiles and Cohen's *d* by reduced ethnic group classification. 75
Table 4. Linear regression of ethnicity reclassified as White-Non-white predicting total SPM raw score 77
Table 5. Linear regression of broad ethnic group predicting total SPM raw score 77
Table 6. Linear regression of reduced ethnic group classification predicting total SPM raw score 77
Table 7. Initial iteration for LR Method identification of White-Non-white ethnic DIF. 81
Figure 13. Test Characteristic Curves based on LR DIF parameters for White (Referent) and Non-white (Focal) groups responding to items of the SPM. 82
Table 8. Initial iteration for LR Method identification of White-Asian ethnic DIF. 84
Figure 14. Test Characteristic Curves based on LR DIF parameters for White (Referent) and Asian (Focal) groups responding to items of the SPM. 85
Table 9. Initial iteration for LR Method identification of White-Middle Eastern ethnic DIF. 86
Figure 15. Test Characteristic Curves based on LR DIF parameters for White (Referent) and Middle Eastern (Focal) groups responding to items of the SPM. 87
Figure 16. Scree Plot of eigenvalues for factors underlying the SPM. 91
Table 10. Fit statistics generated by Mplus for the 1-, 2-, 3- and 4-class models. 93
Table 11. MLVM parameters generated for the SPM (2-class solution). 94
Table 12. MLVM parameters generated for the SPM (4-class solution). 96

Figure 17. TCCs for each latent class in the 2-class model and for the total sample (One-class model). 98
Table 13. Number, mean raw score and z-score for each class within the 2-class model. ... 98
Table 14. Frequency of White and Non-white participants within each latent class. 99
Table 15. Distribution of broad ethnic groups within each latent class. 99
Table 16. Distribution of reduced ethnic categories within each latent class. 99
Table 17. Logistic regression for White-Non-white classification predicting latent class in the 2-class model. 100
Table 18. Logistic regression for broad ethnic classification (Latin-American excluded) predicting latent class in the 2-class model. 101
Table 19. Logistic regression for reduced ethnic classification predicting latent class in the 2-class model. 101
Table 20. Odds ratios comparing odds for each latent class of a correct response to each item in the SPM. Note: Odds ratios greater than 1 indicate higher odds of a correct response for the first class in each comparison. 103
Table 21. Rotated factor matrix for the 2-factor solution underlying the SPM. Note: Factor loadings below .20 have been suppressed. 106
Table 22. Fit statistics generated using a reduced set of SPM items based on loadings on the first extracted factor. 107
Table 23. Fit statistics generated using a reduced set of SPM items based on loadings on the second extracted factor. 108
Table 24. Mean raw score, percentiles and Cohen's *d* by ethnicity classified as White/Non-white. 111
Table 25. Mean raw score, percentiles and Cohen's *d* by broad ethnic group classification. 111
Table 26. Mean raw score, percentiles and Cohen's *d* by reduced ethnic group classification. 111
Table 27. Linear regression of ethnicity reclassified as White-Non-white predicting total APM raw score 112
Table 28. Linear regression of broad ethnic group predicting total APM raw score. 112
Table 29. Linear regression of reduced ethnic group classification predicting total APM raw score 113
Table 30. Initial iteration for LR Method identification of White-Non-white ethnic DIF in the APM. 114
Figure 18. Test Characteristic Curves based on LR DIF parameters for White (Referent) and Non-white (Focal) groups responding to items of the APM. 115
Table 31. Initial iteration for LR Method identification of White-Asian ethnic DIF in the APM. 116
Figure 19. Test Characteristic Curves based on LR DIF parameters for White (Referent) and Asian (Focal) groups responding to items of the APM. 117
Figure 20. Screen Plot of eigenvalues for factors underlying the APM. 118
Table 32. Fit statistics generated by Mplus for the 1-, 2-, 3- and 4-class models. 119
Table 33. MLVM parameters generated for the APM (3-class solution). 120
Figure 21. TCCs for each latent class in the 3-class model and for the total sample (One-class model). 121
Table 34. Number, mean raw score and z-score for each class within the 3-class and 4-class models. 121
Table 35. Frequency of White and Non-white participants within each latent class. 122

Table 36. Distribution of broad ethnic groups within each class. 122
Table 37. Distribution of reduced ethnic categories within each class. 122
Table 38. Multinomial logistic regression for White-Non-white classification predicting latent class in the 3-class model. 123
Table 39. Multinomial logistic regression for broad ethnic classification (Latin-American and Black excluded) predicting latent class in the 3-class model. 123
Table 40. Multinomial logistic regression for reduced ethnic classification predicting latent class in the 3-class model. 124
Table 41. Odds ratios comparing odds for each latent class of a correct response to each item in the APM. Note: Odds ratios greater than 1 indicate higher odds of a correct response for the first class in each comparison. 125
Table 42. Rotated factor matrix for the 2-factor solution underlying the APM. 127
Table 43. Fit statistics generated using a reduced set of APM items based on loadings on the first extracted factor. 128
Table 44. Fit statistics generated using a reduced set of APM items based on loadings on the second extracted factor. 129
Figure 22. Example item from the PfS Abstract (used with kind permission from Team Focus Ltd.). 151
Figure 23. Example item from the Memory and Attention Test. 153
Table 45. Summary of the 13 Trait scales (adapted from ABA, 2011). 156
Figure 24. Scree plot based on eigenvalues of extracted factors underlying the test familiarity scale (pilot study). 163
Figure 25. Scree plot based on eigenvalues of extracted factors underlying the test familiarity scale (full sample). 164
Table 44. Mean raw scores and Cohen's *d* by ethnicity classified as White/Non-white. 175
Table 45. Mean raw scores and Cohen's d by broad ethnic group. 175
Table 46. Linear regression of ethnicity reclassified as White-Non-white predicting total PfS Verbal raw score. 177
Table 47. Linear regression of broad ethnic group predicting total PfS Verbal raw score. 177
Table 48. Linear regression of ethnicity reclassified as White-Non-white predicting total PfS Numerical raw score. 177
Table 49. Linear regression of broad ethnic group predicting total PfS Numerical raw score. 177
Table 50. Linear regression of ethnicity reclassified as White-Non-white predicting total PfS Abstract raw score. 177
Table 51. Linear regression of broad ethnic group predicting total PfS Abstract raw score. 177
Table 52. Linear regression of ethnicity reclassified as White-Non-white predicting total Raven's SPM raw score. 178
Table 53. Linear regression of broad ethnic group predicting total Raven's SPM raw score. 178
Table 54. Linear regression of ethnicity reclassified as White-Non-white predicting MAT Accuracy. 178
Table 55. Linear regression of broad ethnic group predicting MAT Accuracy. 178
Table 56. Linear regression of ethnicity reclassified as White-Non-white predicting MAT Decision Efficiency. 178
Table 57. Linear regression of broad ethnic group predicting MAT Decision Efficiency. 178

Table 58. Linear regression of ethnicity reclassified as White-Non-white predicting MAT Speed of Working. 179
Table 59. Linear regression of broad ethnic group predicting MAT Speed of Working. 179
Table 60. Matrix of Pearson correlations between measures of ability and socio-economic variables. 181
Table 61. R^2, B coefficients and β weights for simple linear regressions of test familiarity predicting total test score. 182
Table 62. Matrix of Pearson correlations between test familiarity and key variables in the dataset. 183
Table 63. Analyses of test familiarity as a mediator of the relationship between socio-environmental variables and test performance. 185
Table 64. Matrix of Pearson correlations between scores on ability measures. 187
Figure 26. CFA model of ability measures made to load on a single latent factor, *g*. 189
Table 65. Matrix of Pearson correlations between ability test scores and Trait personality scales. 191
Table 66. Matrix of Pearson correlations between personality traits and other key variables in the study. 193
Table 67. Mean raw scores and Cohen's *d* on key variables by ethnicity classified as White/Non-white. 195
Table 68. Mean raw scores and Cohen's *d* on key variables by broad ethnic group. 196
Table 69. Summary of the Multiple *R* between mean raw score and ethnicity (White-Non-white classification) that can be explained by test familiarity. 199
Table 70. Summary of the Multiple *R* between mean raw score and ethnicity (broad classification) that can be explained by test familiarity. 199
Table 71. Summary of the Multiple *R* between mean raw score and ethnicity (White-Non-white classification) that can be explained by Optimism. 201
Table 72. Summary of the Multiple *R* between mean raw score and ethnicity (broad classification) that can be explained by Optimism. 201
Table 73. Summary of the Multiple *R* between mean raw score and ethnicity (White-Non-white classification) that can be explained by test performance facilitating and debilitating factors 203
Table 74. Summary of the Multiple *R* between mean raw score and ethnicity (broad classification) that can be explained by test performance facilitating and debilitating factors. 203
Figure 27. Proposed model showing how, in low-stakes situations, ethnicity might give rise to differences in test performance through the mediating links of socio-economic factors, culture, optimism, and test familiarity. 222

CHAPTER 1: INTRODUCTION

Why do some ethnic groups consistently perform more poorly than others on measures of cognitive ability when used for job selection? A definitive answer to this fundamental question has eluded selection assessment researchers for nearly a hundred years. If we could properly understand the nature of ethnic group test performance by understanding its antecedents, we could address a potential barrier to employment for many community groups in the UK.

The present research aims to explore this phenomenon by focusing on the identification of non-cognitive factors that are responsible for ethnic group performance differences, and – ultimately – the root causes of these factors. The position of the research is that there are currently two major gaps in the ethnic test impact/bias literature that cloud our understanding of ethnic group differences. The first is that there is little consensus as to the proportion of ethnic group test score differences that can be accounted for by Differential Test Functioning (DTF). The second is that the root causes of – and the mechanisms by which they give rise to – ethnic group performance differences are still poorly understood.

The present research argues that ethnic group differences are fundamentally socio-environmental in nature. Deeply-ingrained societal differences in the access to education and prosperity afforded to different ethnic groups gives rise to performance differences through key performance facilitating and debilitating factors such as test familiarity and certain personality traits. The implication of this is that these non-cognitive factors can potentially explain much more of the difference in test performance between ethnic groups than has previously been possible. If this is shown to be the case, it would have wide-reaching implications for how we view cognitive differences between ethnic groups and, consequently, how group test score differences are managed fairly in organisations.

This research is particularly relevant for practitioners and policy makers in organisations, employment law, and Human Resource Management, for a number of reasons. In recent history there have been significant changes in the make-up of workforce in the UK. Census data shows that, in the wider population of England and Wales, ethnic minorities have risen from making up 5.9% of the population in 1991 to 7.9% in 2001, and once again to 14.0% when data was last collected for the 2011 census (ONS, 2012). As this trend continues in the future (as it is likely to do), it will become increasingly more salient for organisations to ensure that their selection processes are not disadvantaging minority ethnic groups.

Parallel to this, there has been a gradual increase in the prevalence of organisations in the UK that use ability testing as part of their selection processes. In their most recent annual survey of UK recruitment practices, the Chartered Institute of Personnel and Development estimated that 45% of UK organisations have now adopted general mental ability testing for employee selection, a figure that has doubled since 2011 (CIPD, 2013). However, even after the passing of new legislation in the form of the Equality Act of 2010, there is still no clear legal guidance on either what constitutes bias in selection or how it should be addressed in UK organisations.

Finally, previous research has shown that the degree to which employees perceive an organisation's processes to be fair directly impacts of a number of factors such as their organisational commitment, job satisfaction and overall performance (Johnson et al., 2009). The present research's findings will help organisations to ensure that their selection processes are seen to be treating individuals from all ethnic backgrounds equally, allowing them to get the best from their employees, be they existing or future job incumbents.

1.1 The Present Research

To address the two knowledge gaps outlined in the previous section, two studies were designed and conducted. The focus of both studies was on the practical implications of

ethnic group performance differences. By systematically examining the factors that affect test performance between ethnic groups, this research offers the potential for important real-world impact on how practitioners deploy ability testing in organisations. The following sections will describe the two studies that make up this research, highlighting the novel contribution that they will make to both academia and practice.

1.1.1 Study 1

The first study in this research will focus on whether ethnic group test performance differences exist in a new, up-to-date sample of global test data, and – if this is found to be the case – to what extent these differences can be explained by Differential Test Functioning. The study will employ two competing statistical techniques for the identification of DTF within two large archival samples of test data. This first, the Logistic Regression (LR) Method, is a traditional technique for the identification of Differential Item Functioning (DIF) – and, by extension, DTF – recommended by French and Maller (2007) that is useful for the identification of both uniform and non-uniform ethnic DIF. The second technique to be utilised will be Mixture Latent Variable Modelling (MLVM), the latent classes generated by which should give some insight into the underlying factors that give rise to ethnic DTF. Combining these two techniques – both of which are viewed as robust approaches to the identification of the precursors of measurement bias – represents the best chance for definitively addressing whether DTF is a productive line for ethnic bias researchers to be pursuing, or whether it is, ultimately, unable to explain more than a trivial proportion of the differences in test performance between ethnic groups.

The archives in this study were chosen for a number of reasons. Firstly, the archives contain test data from the Raven's Standard Progressive Matrices (SPM) and Advanced Progressive Matrices (APM) respectively, measures that have both been demonstrated consistently to show large mean test score differences between ethnic groups, most frequently between

Black and White test takers (e.g. Jensen, 1998; Owen, 1992). Secondly, they represent new, unpublished data that will give insight into the current state of ethnic group differences in test performance. Much of the published research on ethnic test performance differences is based on data that is now outdated. This is particularly relevant in the DIF/DTF literature, as the large sample sizes required for adequate power when employing these statistical techniques often means falling back on data archives containing test data that was collected a long time ago, meaning that, at best, they provide a snapshot of the state of ethnic DTF from the recent past. Thirdly, the archives contain test data from a global sample of candidates. It was reasoned that the global nature of the archives would allow findings to be generated that would be generalisable across national borders, considering ethnic differences on a much larger scale than is typical in the often US-centric literature.

The first study addresses an important gap in the current literature on ethnic test impact/bias. To this point, there has been no clear consensus on the extent to which DTF plays a part in ethnic group test score differences. This is due in no small part to the methodological problems that have dogged this part of the field for many years. A robust investigation of ethnic measurement bias will clarify the literature. If it can be demonstrated that DTF can effectively explain much of the difference between ethnic groups' test performance, it has important implications for how we – as both practitioners and academics – view ability testing as a legitimate method by which to select candidates in multicultural societies.

1.1.2 Study 2

The second study aims to address the issue of identifying the root causes of ethnic test score differences, and the mechanisms by which these factors influence performance. While the evidence in the literature would suggest that ethnic group differences in performance are much more strongly attributable to differences in socio-environmental factors (e.g. Dickens & Flynn, 2006) than through differences in heritable factors (e.g. Rushton & Jensen, 2005),

little is understood of the mechanism by which environmental factors lead to differences in performance. This represents a substantial gap in the literature, as this lack of understanding makes it difficult to address ethnic differences in test performance in a sensible way: Without knowing the nature of the impact that these factors have on the test scores of members of these groups, any intervention to ameliorate them must be based on speculation.

To properly test the study's theoretical model (see section 2.5.3), a sophisticated approach to data collection needed to be employed. On the basis of the model, a second, primary data study was designed. The study set out to capture a wide range of demographic and socio-environmental variables, with the aim of examining the root causes of ethnic group performance differences and the mechanisms by which they arise. The study's design imposed tight controls on the environment in which test data was collected, both at the microscopic level (in terms of the test environment) and the macroscopic level (in terms of the national setting to which the research pertains).

To investigate the influence of *g*-saturation on ethnic group differences, with the aim of better understanding the Spearman-Jensen Effect, test data from a wide range of ability measures was collected, including measures designed to assess specific abilities such as verbal, numerical and abstract reasoning, but also constructs related to general mental ability such as attention and working memory capacity (WMC). This allowed for the effect of test familiarity as a mediator of the relationship between ethnicity and test performance to be examined across tests designed to measure different constructs, with different *g*-loadings, and different balances in the proportions of their scores that could be attributable to g_f and g_c. All tests were administered in controlled conditions, to simulate the proctored environment in which candidates would normally complete selection tests during late-stage assessment, albeit outside of a high-stakes situation. This allowed test conditions to be

relatively consistent, minimising the random error variance in test scores that could be introduced through variations in noise, light, heat, interruptions and so on.

In addition to these ability measures, a robust trait measure of personality was also administered to participants. This measure, the Trait Personality Inventory (Trait; ABA, 2011), has been designed specifically for use in organisational settings for the purposes of selection and development. As such, it contains scales that measure facets of all of the Big Five personality factors, as well as other personality constructs that are relevant to performance in a variety of occupational settings. Measuring participants' personality in this way allowed for the investigation of these traits as test performance facilitating and debilitating factors that could potentially explain variance in test scores between ethnic groups over and above that explained by test familiarity.

The findings of Study 2 will allow both academics and practitioners alike to better understand the nature of test performance differences between ethnic groups. In doing so, it will address an important gap in the literature that will ultimately allow ability testing to be deployed in a robust way that is least likely to disadvantage members of minority social groups.

CHAPTER 2: LITERATURE REVIEW

This chapter will review the extant literature in a number of key areas relevant to ethnic group differences in cognitive test performance. It will begin by discussing how researchers typically understand the underlying structure of intelligence and how ability testing can be deployed to provide a measure of these hidden constructs and predict future performance in selection. It will go on to explore the key limitation of using ability tests for selection, namely that of persistent ethnic group performance differences. Finally, two key knowledge gaps within the ethnic test/impact bias literature will be identified and explored.

2.1 The Structure and Measurement of Cognitive Ability

A number of cognitive factors have, since the 1900's, been generally accepted as underlying intelligence (Robertson & Smith, 2001). These factors are known by different names, but they are most commonly known as fluid intelligence, crystallised intelligence, visualisation, retrieval, and cognitive speed. Modern ability tests are still based upon models of intelligence that incorporate these factors (Robertson & Smith, 2001).

One such model is Carroll's (1993) three-stratum hierarchical structure of cognitive ability, shown in Figure 1 below.

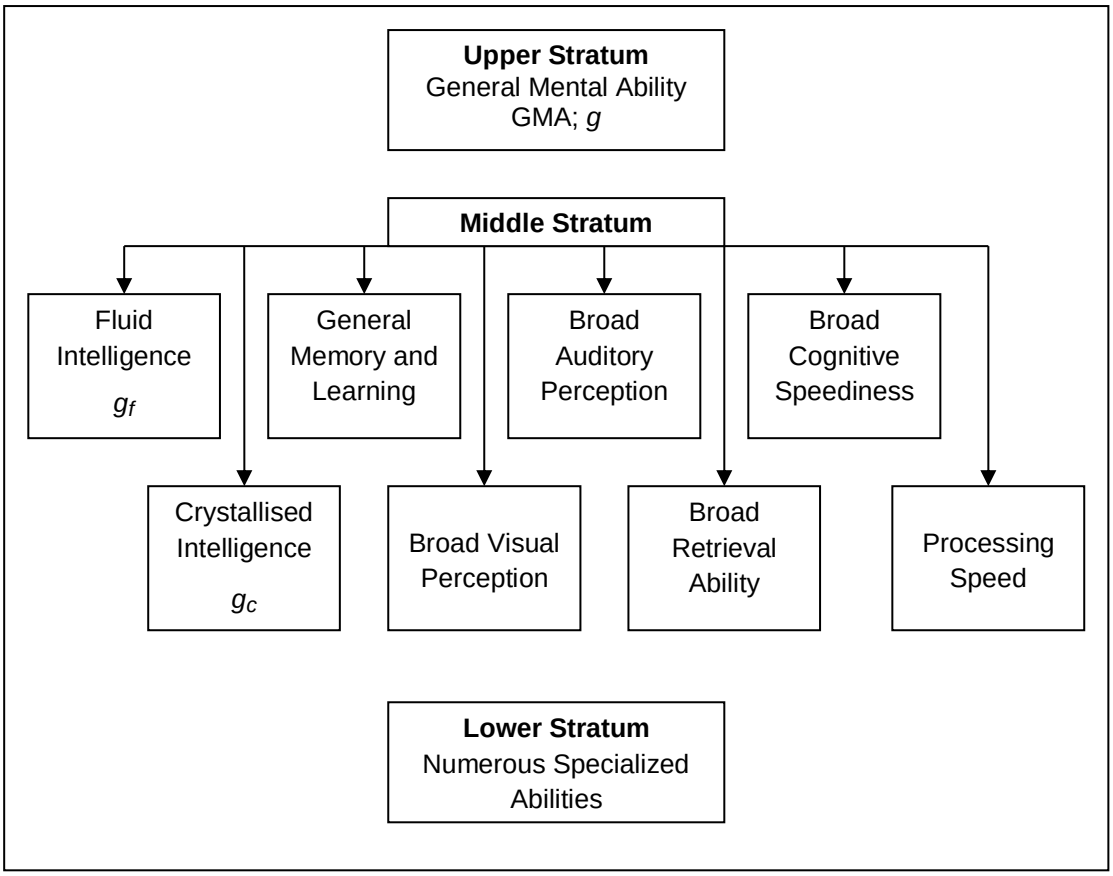

Figure 1. Carroll's (1993) three-level hierarchy of intelligence (adapted from Robertson et al, 2002)

At the top of the hierarchy of intelligence is general mental ability (known, variously, as GMA or *g*), a general factor of intelligence first proposed by Spearman (1927). *g* is an indication of a person's ability to process, store and retrieve information, and is statistically related to all factors lower down in Carroll's hierarchy. At the middle stratum, the major facets of general intelligence are sub-divided into eight distinct types of intelligence. At the lowest of the three strata, intelligence is further sub-divided into a large number of specific facets of intelligence, such as verbal, numerical and abstract reasoning, and spatial rotation.

Though there has been much debate of the underlying structure of intelligence in the literature, most models, such as Carrol's (1993) or those of Cattell (1963) and Horn (1976), agree that *g* itself can be subdivided into two factors, g_f and g_c. g_f refers to fluid

intelligence, also classified as “intelligence-as-process” by Ackerman (1996), and “mechanics of the mind” by Baltes (1990; cited by Warr, 2001). Fluid intelligence is a person’s ability to think and reason in abstract ways. This contrasts with crystallised intelligence (g_c), which relates to a person’s knowledge and skills that have been acquired through both formal and informal learning, so is comparatively much more experience-based. Though all Lower Stratum facets of ability tap into both g_f and g_c to some degree, the extent to which this is the case is non-uniform across specific abilities. For example, Beauducel, Brocke and Liepmann (2001) have argued that verbal reasoning has a much larger component related to g_c than it does to g_f, whereas abstract reasoning (sometimes referred to as figural reasoning) tends to display the opposite pattern.

A key issue in the measurement of *g* and the factors into which it is most frequently subdivided is that they represent hidden constructs that cannot be directly measured. Instead, attempts to identify the structure by which these cognitive constructs are arranged typically make use of statistical techniques such as latent variable modelling. These techniques allow researchers to draw inferences about a person’s true level of ability on the basis of some overt, measurable predictor (or predictors). In the case of the reasoning ability represented by fluid intelligence, research has consistently shown that its strongest predictor by far is working memory capacity (WMC; e.g. Conway et al., 2002; Engle et al., 1999). However, the use of WMC measures to estimate a person’s true level of ability is limited outside of experimental psychology settings. A more practically-applicable approach to measuring cognitive ability comes in the form of standardised ability tests.

Ability tests provide a standardised sample of cognitive ability that can be described using a numerical scale, measuring a participant’s performance in a single dimension of ability such as verbal, numerical or abstract reasoning by making inferences about their true ability in these areas based on the number of items relating to a specific area they answer correctly (Cronbach, 1990). Ability tests are appropriate for assessing ability in a variety of

environments. They may be used to get a measure of a participant's ability in clinical, educational and organisational settings cheaply and efficiently (Arnold et al., 2005).

2.2 The Relationship of Intelligence to Job Performance

In the case of ability tests' use in organisational settings, they are most frequently deployed in job selection processes to identify the strongest candidate from a pool (known as selecting in), or to screen out candidates who fail to meet a minimum threshold of ability (as part of a second sift, sometimes referred to as selecting out). Though not immediately apparent, the strength of ability testing in selection lies in its validity as a predictor of important organisational criteria, most frequently future job performance. Meta-analytic research into different selection methods has consistently shown that ability tests are the single best predictor of future job performance available (e.g. Schmidt & Hunter, 1998; Robertson & Smith, 2001). The predictive validity of a selection method is a measure of how accurately the measure predicts future job performance and is usually expressed as a correlation coefficient of between 0 (indicating no predictive power) and 1 (indicating perfect prediction of the given criterion). Though there is some variability in reported figures, it is generally accepted that the mean correlation coefficient for the relationship between ability test score at selection and job performance lies between .5 and .6 (Bertua, Anderson & Salgado, 2005), a very high figure for an individual selection method. Moreover, these findings appear to be generalisable across a variety of occupations.

There are a number of explanations for these observations, all of which hinge on the assumption that cognitive ability tests tap in, to some extent, to general mental ability. Firstly, ability tests may assess general competencies that are applicable to a variety of jobs. Certainly, with the increased cognitive demand (for example, with the increase in technological complexity) in the vast majority of occupational settings today, cognitive ability may be an increasingly important factor in predicting job success (Millward, 2005). More

specifically, it has been suggested that a high level of *g* aids the acquisition of declarative and procedural job knowledge (e.g. Kuncel, Ones & Sackett, 2010), allowing those with a high level of *g* to attain a desirable level of performance more quickly than their peers of lower *g*. Consistent with this hypothesis, several studies have shown cognitive ability to be a valid predictor of training success (e.g. Ackerman, Kanfer & Goff, 1995; Ree & Carretta, 1998; Ree & Earles, 1991).

2.3 Ethnic Group Differences in Ability Test Performance

In spite of the clear advantage of using ability tests to predict future job performance over more traditional methods (such as unstructured interviews), academics have often warned that there are issues associated with their deployment in the high-stakes decision making process of job selection. The most frequently cited drawback of using ability tests for selection in the literature is that of ethnic bias. It has often been observed that certain ethnic groups score less well on measures of ability than the White majority (e.g. Cooper & Robertson, 1995). Far and away the most frequent-cited ethnic differences in test performance are those between White and Black test takers, in particular those from the US. In a review of the extant literature, Ruston and Jensen (2005) observed a mean difference in test scores of 1.1 standard deviation (S.D.) units between Black and White test takers. The work of Gottfredson (2005) corroborates this observation, reporting Black-White differences of between 0.8 and 1.2 S.D. units. These observations, however, have not been limited to the case of Black-White test takers. Herrnstein and Murray (1994) observed Hispanic-White differences in performance on IQ measures to amount to around 14 IQ points, a difference in S.D. units of slightly less than 1 (given that the standard deviation of the IQ scale for most measures is 15). Findings pertaining to differences between Asian ethnic subpopulations (e.g. Middle Eastern, East Asian, South East Asian) and White test takers are often more difficult to interpret, due in no small part to the habit within the literature of simplifying these distinct groups to a single ethnic group (‘Asian’). There is, however, some evidence to

suggest that East Asian test takers perform comparably well, if not marginally better, than their White counterparts (Rushton & Jensen, 2005). When ethnicity is considered at its broadest level of abstraction (i.e. when it is reduced to a White/non-White dichotomy), there is evidence to suggest that there are test score mean group differences of somewhere between 0.46 (Martocchio & Whitener, 1992) and 0.83 S.D. units (Schmitt, Clause & Pulakos, 1996) between White and non-White test takers.

What is perhaps more concerning is that these group differences are typically more pronounced than they are for alternative selection methods, those that often display lower predictive validity. Outtz (2002) showed that ability tests tend to produce differences in scores between ethnic groups that are on average 3 to 5 times larger than for personality tests or structured interviews. Similarly, Dean, Roth and Bobko (2008) found meta-analytic evidence of ethnic differences in assessment centre ratings of 0.52 S.D. units favouring White people over Black people, and differences of 0.28 S.D. units favouring White people over Hispanic people. In a meta-analysis of ethnic group differences in situational judgement test (SJT) performance, Whetzel, McDaniel and Nguyen (2008) observed Black-White differences of 0.38 S.D. units, Hispanic-White differences of 0.24 S.D. units, and Asian-White differences of 0.29 S.D. units. The implication, here, is that selection assessment practitioners are frequently forced to make decisions on the most appropriate methods to deploy based on trade-offs between predictive validity, and fair and equitable treatment of minority ethnic groups in their selection processes.

While the existence of these test score group differences is well established, there is much disagreement in the field as to their fundamental nature. Specifically, two key issues currently face the ethnic test performance difference literature as it stands. Firstly, there is disagreement over whether or not tests function in a similar fashion for candidates of all ethnicities (i.e. whether they accurately measure ability in the same way across ethnic groups). Secondly, there is very little consensus as to what principle factors might cause people of one ethnicity to perform better on these tests than those of another.

Rather than being separate, these two issues are necessarily interdependent. Both relate – at their core – to the wider issue of *why* ethnic test performance differences occur, and – by extension – how best to manage them in organisations. The following sections will address these two issues in turn to establish the present state of the understanding of ethnic group differences in the extant literature.

2.4 Competing Conceptualisations of Group Differences: Test Impact and Test Bias

Though the ethnic group test performance differences described above are well documented in the literature, it is important to recognise that these differences are not necessarily representative of some kind of systematic bias in and of themselves. It is entirely possible that group test score differences are so because they represent true differences in cognitive ability between those groups. In situations in which only group differences are observed, but a true difference (or lack thereof) between these groups in their true level of ability cannot be inferred, a term more appropriate than test bias is test *impact*. Test impact is defined as a measure of one group's test performance relative to another, most frequently expressed in standard deviation units (Vogt, 1999, cited by Zwick, 2004). At its core, then, it is purely concerned with observable performance differences. Conversely, the concept of bias, broadly speaking, implies that one group is systematically treated more favourably than another in some way, either as a *consequence* of these performance differences, or as a *precursor* to them.

Millsap (2007) takes this a step further, arguing that 'bias' as it is conceptualised in much of the selection literature is a misleading term. He maintains that it is of critical importance that researchers are clear and transparent about the particular conceptualisation of bias with which they are concerned, as different types of invariance (i.e. lack of bias) have been demonstrated to be potentially mutually exclusive under realistic conditions (see section 2.4.4.3 for more detailed description of this phenomenon).

The upshot of all of this, then, is that crucial distinctions need to be made between differences in test performance and the variety of variables with which they may be associated. Ethnic group differences can manifest in a number of different ways, potentially leading to differential outcomes between ethnic groups in terms of different group selection ratios (referred to as Adverse Impact; e.g. Bobko & Roth, 2009), systematic under- or overestimation of a group's future job performance (violations of the Cleary Rule; Cleary, 1968), or overestimation of the magnitude of subsequent performance differences between groups (violations of the Thorndike Rule; Thorndike, 1971, cited by Chung-Yan & Cronshaw, 2002). In addition to these consequences of group difference, differences in the way in which a latent trait of interest is estimated by the test (e.g. Millsap, 2007) can give rise to test performance differences at the group level. Fundamentally, these distinctions are at the core of a key issue facing the field, one for which there has been little consensus between researchers, namely the question of what constitutes ethnic bias on a conceptual level.

2.4.1 Adverse Impact

The most frequently cited definition of the circumstances under which ethnic group performance differences at selection represent bias is when a measure is demonstrated to have adverse impact on a particular group. In the US legal system, a selection measure is judged to produce adverse impact on a specific group if it is found to be in violation of the 80% Rule, often known as the 4/5^{ths} Rule (Equal Employment Opportunity Commission). According to this rule of thumb, if the selection ratio of a minority group is found to be smaller than 80% of the selection ratio for the higher-scoring majority group in a selection process, that process can be judged to have adverse impact on the minority group (Bobko & Roth, 2009), though it is worth noting that this is heuristic rather than law *per se*. A graphical representation of the real impact of using a measure that displays adverse impact on an ethnically diverse candidate pool in a selection process is shown in Figure 2 below:

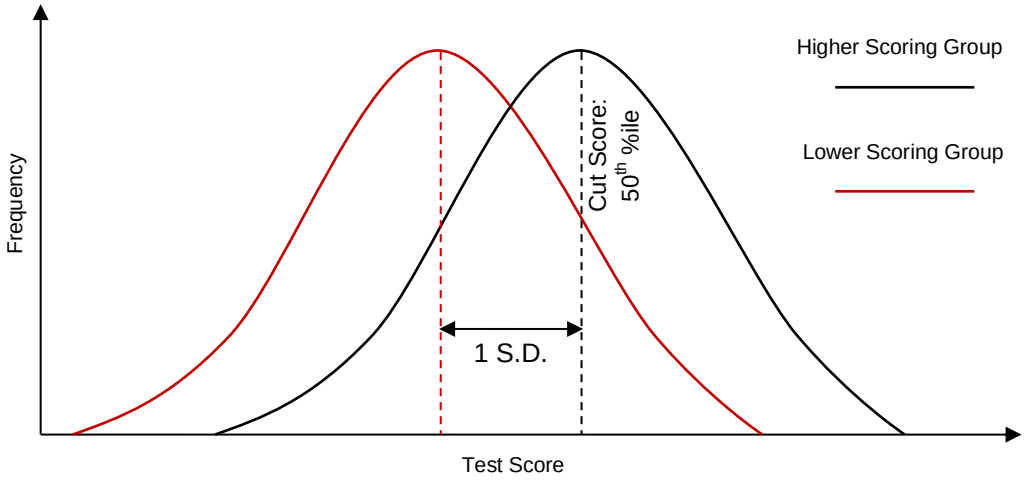

Figure 2. The practical impact of a 1 S.D. difference in mean test score between groups at selection.

The most attractive quality of the 80% Rule is that it articulates, in simple terms, the real world effect of what can be considered – at least statistically – to be relatively small group differences. Assuming, as in Figure 2 above, a normal distribution of test scores for both higher- and lower-scoring test taker groups, and a mean difference in test score between the groups of around 1 S.D. unit, if the cut score for a test is set at the 50th percentile of the higher-scoring group (a relatively conservative cut score, and one that might reasonably be deployed early on in the selection process, perhaps during a second sift), proportionally many more of the lower-scoring group would be selected out of the process by the test than of the higher-scoring group.

While this definition of ethnic bias is useful in practical terms in that it gives clear guidelines for the identification of bias, it is not without its limitations. In the legal context, this definition of bias is only present in US law: UK and European law only acknowledge that a selection procedure that displays adverse impact could potentially be deemed discriminatory and that selection methods that have an 'unfavourable effect' on a specific social group should be examined against technical validity criteria to assess their appropriateness (Higuera, 2001). This legal position makes any attempt to defend or prosecute users of selection tools largely dependent on subjective judgements. A second, and more fundamental, problem with this

rule was noticed by Roth et al. (2006). They observed that its use generates false positives (i.e. the detection of adverse impact when there are not significant between-group differences in mean test scores) relatively frequently.

2.4.2 Cleary's Rule: Differential Prediction

Due to these criticisms of adverse impact as a conceptualisation of bias, many academics tend to discount it in favour of one that is somewhat more sophisticated. Cleary's (1968) model states that a test is biased if it consistently over- or underestimates a certain group's future job performance based on their test performance. Figure 3 shows an example of a violation of Cleary's Rule:

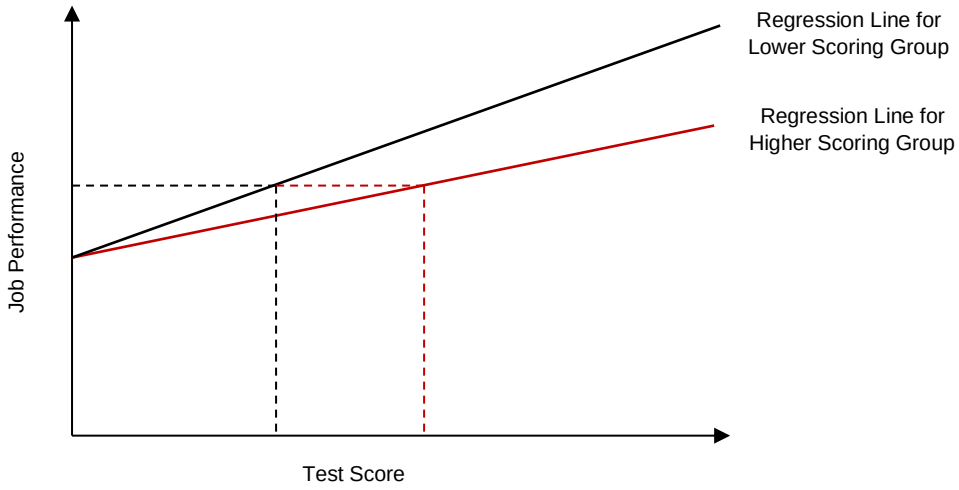

Figure 3. Graph showing different regression lines for higher and lower scoring groups on the same measure of cognitive ability.

In Figure 3, the two examined groups have regression lines that share an intercept, but differ in their slopes. Consequently, people from the lower scoring group will be predicted as having the same level of future job performance as those in the higher scoring group. This would provide evidence of prediction bias. Aguinis, Culpepper and Pierce (2010) argue that

prediction bias can come about if two groups' regression lines differ either in their slopes, intercepts, or both. Therefore, if both groups were to share a common regression line, with equivalent slopes and intercepts for each group, this would indicate that the test was predicting future performance in an unbiased way for members of each group, irrelevant of their differences on the manifest characteristic of interest (for example, ethnicity). For both academics and practitioners, this is an attractive definition of bias with which to work, as it recognises that test scores in selection must necessarily be linked to the criterion that they are being used to predict. This puts it at a clear advantage to the 80% Rule, as the latter might, plausibly, label a predictor that showed ethnic group differences as biased when, in actual fact, these differences were reflective of real performance differences between groups.

For these reasons, Cleary's is a definition commonly applied to ability test scores when bias is discussed in the literature. Berry, Clark and McClure (2011) found meta-analytic evidence of differential prediction between ethnicities that appear to generalise across very large sample sizes. They observed that, while the mean validity coefficient for Asian samples was approximately equal (mean r = .33), mean coefficients for Black samples (mean r = .24) and Hispanic samples (mean r = .30) were lower than the mean of those for White samples (mean r = .33). As a measure of bias, however, Cleary's Rule has been extensively criticised in that it might miss some of the more nuanced differences in outcomes between groups under certain circumstances. One such criticism has led directly to an alternative conceptualisation of bias in the form of Thorndike's Rule.

2.4.3 Thorndike's Rule

Chung-Yan and Cronshaw (2002) proposed a new definition of bias based on Thorndike's (1971, cited by Chung-Yan & Cronshaw, 2002) model of fairness. Thorndike's original argument was that a test could show evidence of 'unfairness' (a relatively subjective

judgement of the unsuitability of a test) even if it had been judged to be free from 'bias' (the identification of sources of systematic variance based on statistical analysis) according to Cleary's Rule. Chung-Yan and Cronshaw (2002) revised Thorndike's model for the use of bias detection rather than fairness as it had been originally intended, arguing that a test could still show evidence of bias even if it was not in violation of Cleary's Rule, as in Figure 4 below.

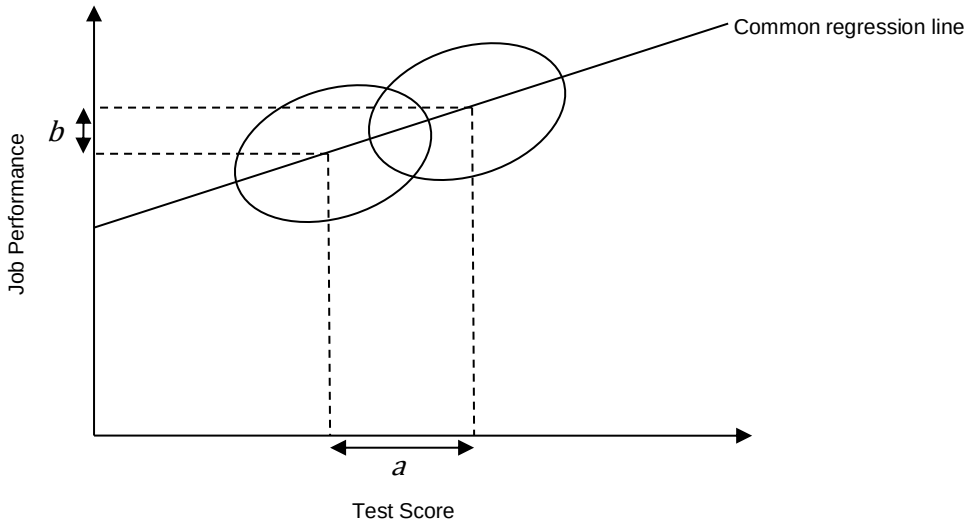

Figure 4. Graph showing disparity in size of mean group difference in job performance and test score for two separate social groups.

In Figure 4, different groups share a common regression line so the test does not display differential prediction for the two groups (i.e. validity coefficients should be approximately equal). However, because the difference between group mean test scores at selection (*a*) is markedly bigger than the difference between group mean future job performance (*b*), Chung-Yan and Cronshaw argue that the test in this example shows evidence of selection bias in that group performance differences would be exaggerated by test score differences. In practical terms, these differences appear to have a degree of empirical support in that meta-analytic evidence of overall performance ratings has shown mean differences between Black and White employees of 0.39 S.D. units favouring White groups (Kraiger & Ford, 1985), a

much smaller degree of difference than has been demonstrated in mean test scores for the same ethnic groups (e.g. Rushton & Jensen, 2005).

However, researchers have subsequently criticised Thorndike's model of fairness (and – by extension – its reconceptualisation as a model of prediction bias) based on a number of hidden criteria that need to be satisfied in order for the rule to be applied effectively in all cases. The most important of these, Reeve and Bonaccio (2009) argue, is that Thorndike's model assumes that the criterion of subsequent performance used is 100% reliable. However, research into performance measurement has found that the majority of performance measures tend to be unreliable. In particular, the issue of inter-rater reliability in subjective performance rating systems is of particular importance given the prevalence of this approach to measuring job performance (Viswesvaran, Ones & Schmidt, 1996). These observations point to the conclusion that the revised Thorndike rule as a model of bias is limited in its practical application.

The message, therefore, is that there is very little consensus in either the literature or in legislature as to what constitutes a good working definition of bias that can be applied reasonably to all selection processes and in all situations. There is, however, a single thread that runs through all these definitions, an assumption that underlies each of them, whether stated explicitly or implicit within them, that can be condensed to the following statement: If ethnic group performance differences are not the direct result of true differences in ability as a valid predictor of important organisational criteria, these differences are indicative of ethnic bias.

2.4.4 Measurement Invariance and Item Response Theory

An alternative to models of prediction bias, such as that of Cleary and the adapted Thorndike model, which focus on differential prediction of the outcomes of selection, measurement invariance focuses on whether observed test score differences are attributable to true

difference between groups on the latent trait (or traits) being measured, or to differences in how the test functions for these groups. Millsap (2007) provides a more formal definition of measurement invariance expressed in terms of conditional probability. In this definition, scores on an observed measure (such as an ability test) are represented by X. The latent variable (or variables) measured by X are represented by W. Finally, person characteristics of interest (such as ethnicity) that should be irrelevant to X when accounting for W are represented by V. On the basis of these three vectors, Millsap derives the following equation to represent the conditions under which measurement invariance holds:

$$P(X|W,V) = P(X|W) \quad (1)$$

Therefore, Millsap's definition states that, for measurement invariance to exist, the conditional probability for X (either as a discrete conditional probability for discrete variables represented by X, or as a conditional probability function for continuous ones) given W and V *must* equal the conditional probability of X given W (i.e. irrelevant of V). In practical terms, what a violation of Equation 1 would represent would be unequal distributions of test scores conditional on true ability between groups of interest.

Borsboom (2006a) reasons that fairness and equity in testing cannot exist without measurement invariance, and recommends that an investigation of measurement invariance should be routinely performed in any situation requiring unbiased selection. Gomez-Benito et al. (2010) argue that the existence of a lack of measurement invariance represents a serious threat to the validity of a measure. Conversely, the demonstration of measurement invariance represents evidence for generalisability of measurement across demographically heterogeneous populations (for example, multi-ethnic populations such as the UK). Many modern approaches to the investigation of measurement invariance make use of methods based on Item Response Theory to detect differential item functioning (DIF).

Item Response theory (IRT) was initially developed as an alternative to Classical Test Theory (see section 2.5.1 below). Its central tenet is that the probability that a participant will

respond correctly to a test item is based on a number of parameters, an individual's response to a particular test item depending on the interaction between the characteristics of a test item and their own 'person parameters' (Embretson & Reise, 2000). A number of IRT models exist and each attempts to predict the probability of a correct response to a given item in subtly different ways. The most commonly used IRT models are the 2PL (two-parameter logistic) and 3PL (three-parameter logistic) models. Both models predict the probability ((θ)) of a person, s, correctly responding to an item, i. In the case of ability testing, the parameter b_i represents item difficulty, a_i represents how well the item discriminates between respondents, and θ_s represents a 'person parameter' (i.e. some trait of the respondent). The relationship between these parameters is described by the 2PL equation below (adapted from Embretson & Reise, 2000):

$$p(X_{is} = 1|\theta_s, a_i, b_i) = \frac{e^{a_i(\theta_s - b_i)}}{1+e^{a_i(\theta_s - b_i)}} \quad (2)$$

In the above equation, the probability (p) of respondent s providing a response X that is equal to 1 (i.e. a correct answer to a dichotomous correct/incorrect item) to an item i, given that participant's level of θ and the discrimination (a) and difficulty (b) of that item, is described by a logistic function based on these parameters.

The 3PL model additionally adds a random guessing parameter, c_i, to control for respondents randomly selecting the correct response to a multiple choice item. The probability of a person s responding correctly to item i according to the 3PL model is given in the equation below (adapted from Embretson & Reise, 2000):

$$p(X_{is} = 1|\theta_s, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{e^{a_i(\theta_s - b_i)}}{1+e^{a_i(\theta - b_i)}} \quad (3)$$

In other words, the likelihood that a given person answers a particular question on a test correctly depends on a number of factors. Firstly, the level of a trait within the person (for example, their level of verbal ability) will affect the probability of a correct response. Similarly, the difficulty of the question, how well it discriminates between high-ability and low-

ability people, and how likely it is that someone would respond correctly to it if they guessed entirely at random will all affect the probability of a correct response.

By entering values for each coefficient into the equation, an item characteristic curve (ICC) can be generated for item i. An example of what an item characteristic curve might look like is shown in Figure 5 below.

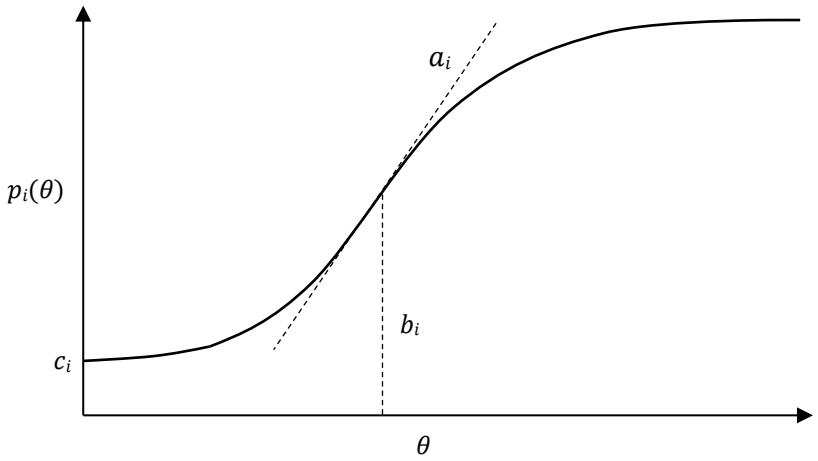

Figure 5. Example ICC for an item according to the 3PL model.

The ICC above shows the probability of a person correctly responding to the item dependent on their level of a trait, θ. A person with low θ will therefore have a lower probability of correctly responding to the item than one with a high level of θ. The form that an item's ICC takes is based directly on its item parameters. Parameter a_i describes the slope of the curve at its steepest point. The item's difficulty, represented by parameter b_i, corresponds to the level of θ a participant would need to possess to have a probability of .5 (i.e. 50% chance) of getting the answer correct (the point on the x-axis at which the curve is steepest. The probability of a person of very low ability randomly guessing the correct answer to the item, parameter c_i, is represented as the lower asymptote of the curve, where it meets the y-axis.

The lesser-known 4PL model includes a fourth parameter, d, used to represent the upper asymptote of the curve (Barton & Lord, 1981). Its function is similar in nature to the d

parameter. Whereas the c parameter is used to estimate the probability of randomly guessing the correct response to an item for participants of a very low level of θ, the d parameter estimates the maximum probability that participants of very *high* levels of θ will have of responding correctly to an item. This upper asymptote reflects the fact that even very high ability participants sometimes make mistakes, so the probability of a participant giving the correct response to an item will never be equal to exactly 1, no matter how easy the item (Liao et al., 2012). Expressed mathematically, the influence of this parameter representing the curve's upper asymptote on the model is to replace the term 1 - c in the 3PL model with d - c. Though d's inclusion in logistic IRT models should, theoretically, improve their accuracy, it is seldom used in practice due to the difficulty in accurately estimating a curve's upper asymptote (Magis, 2013), though it is beginning to see more applied use in the field of computer adaptive testing (e.g. Liao et al, 2012; Yen et al., 2012).

Embretson and Reise (2000) say that the 3PL model is most appropriate for examining dichotomously scored items such as those found in ability tests (i.e. scored as being either 'correct' or 'incorrect'). There are two assumptions that need to be satisfied for the 3PL model to be used for this purpose. Firstly, each item must assess a single construct (e.g. verbal ability). Secondly, a single underlying trait must be responsible for the relationship between the items (i.e. its latent structure should be unifactorial). In the case of a test of specific ability, these assumptions are likely to be met.

2.4.4.1 Differential Item Functioning and Differential Test Functioning

One application of IRT methods that is of particular interest to researchers is in the identification of Differential Item Functioning (DIF). DIF is examined using a collection of statistical methods designed to identify test items that may be potentially problematic in that they behave differently for specific social groups (Zumbo & Gelin, 2005). DIF occurs when test takers of equal true ability have different probabilities of responding to a test item

correctly. If an item shows different ICCs for focal (e.g. minority) and referent (e.g. majority) groups when members of both groups are matched on their level of true ability, this can be taken as evidence of DIF for that item (Zumbo, 2007). An example of DIF is shown in Figure 6 below.

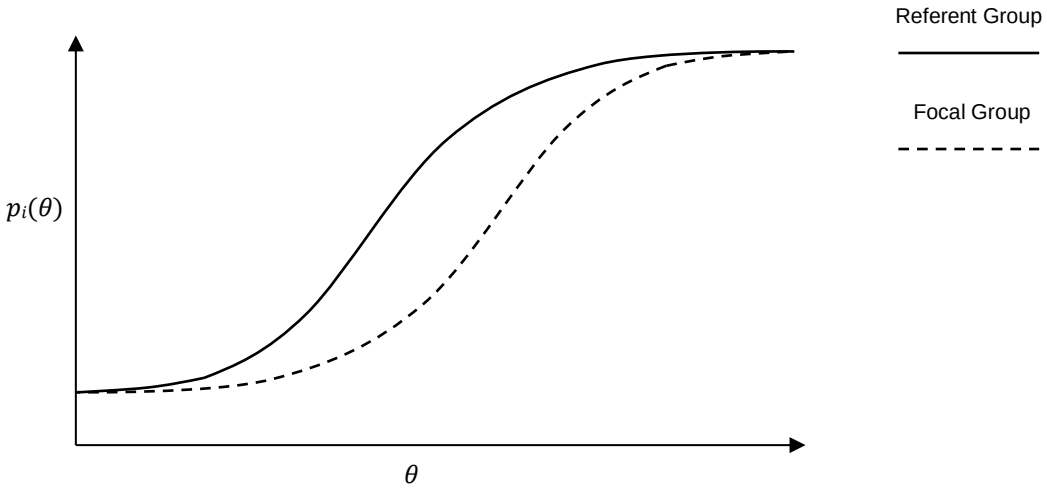

Figure 6. ICCs for an item that shows DIF.

In Figure 6, focal and referent groups display different ICCs for the same test item. The implication of this is that, for any given level of θ, members of focal and referent groups will have different probabilities of answering the test item correctly. Since, at any point on the x-axis, θ should be equal for both groups (i.e. they have the same level of true ability), this indicates that this test item functions differently for different groups of test takers. If the referent group here were White test takers and the focal group were Black test takers, for any given level of true ability (with the exception of extremely low or extremely high levels of ability), a Black test taker would have a lower probability than a White test taker of answering the item correctly, indicating ethnic DIF.

In spite of this, detection of a DIF effect does not, on its own, constitute evidence that a test demonstrates a lack of measurement invariance in favour of either the focal or referent groups, as often DIF effects will combine and cancel out to produce no overall difference in functioning between the two groups at the test level (Gibson & Harvey, 2003). If, however,

when considered together as a whole test, DIF items favour one group over another, this is referred to Differential Test Functioning (DTF). Analogous to Millsap's (2007) definition of measurement invariance described in Equation 1, DTF may be examined by comparison of the conditional probability functions between focal and referent groups. These functions are called Test Characteristic Curves (TCCs), and are generated by summating the ICCs for every item within a test, so take the form of logistic functions. The difference between TCCs for focal and referent groups gives an indication of the impact that group membership has on test scores, conditional on each particular level of θ.

The advantage of employing IRT methods when investigating measurement invariance is clear: They allow researchers to examine whether test score group differences arise as a result of true differences in ability between groups, or whether they represent differences in how a test functions for these groups.

DIF does not, however, always present in the way depicted in Figure 6. Gibson and Harvey (2003) draw the distinction between uniform and non-uniform DIF. Uniform DIF occurs when an item functions less favourably for members of all levels of ability within a particular group (as in Figure 6). In contrast to this, non-uniform DIF occurs when item functioning favours high-ability members of a group while simultaneously favouring low-ability members of the other group, as in Figure 7 below:

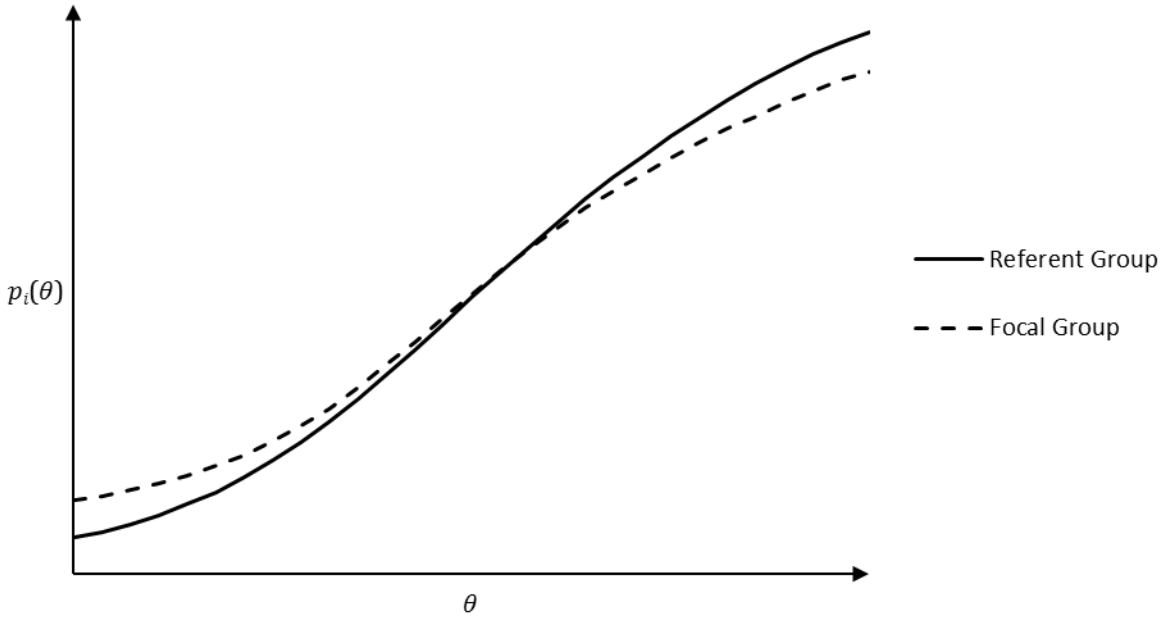

Figure 7. Example ICCs that might be observed for an item that shows Non-uniform DIF.

In Figure 7, focal and referent groups again display different ICCs. As is the case for uniform DIF, members of the referent group of high ability will have a higher probability of responding correctly to the item than members of the focal group of equivalent ability. However, at lower levels of ability, members of focal group have the advantage, having a higher probability of responding to the item correctly than member of the referent group.

Gibson and Harvey observed that uniform and non-uniform items can combine to produce strange effects at the scale level. Their findings were that DTF varied enormously in its nature, some tests favouring the focal group, some the referent group and others still showing no overall DTF favouring one group or another. Borsboom (2006a) highlights this last condition as being particularly problematic for research of this nature. If individual DIF effects are cancelled out in this way, differential functioning tends to occur at the level of conditional population distributions rather than at the individual level. The implication of this is that a measure might still disadvantage individuals in a group when used for selection although research is likely to not detect this effect.

Stark, Chernyshenko and Drasgow (2004) examined DTF in terms of its implications for adverse impact on groups in selection. In their study they found that the overall DTF of a college admission test for US school pupils on explained group differences of around a quarter of a raw score point. Their conclusion was that the overall effect of DTF on adverse impact is negligible. This might lead researchers to conclude that DTF – and, by extension, DIF – is of no use when attempting to explain test impact. However, Cohen and Bolt (2005) have demonstrated that manifest characteristics (such as ethnicity) are often only very weakly related to the latent groups that are being disadvantaged by items that show DIF, as the underlying assumptions of traditional DIF approaches is that focal and referent groups are homogeneous. This could explain why attempts to explain ethnic group test score differences in terms of how items (and, by extension, tests) function for different ethnic groups have only been able to explain somewhere between 0.02 and 0.25 S.D. units of the differences between ethnic groups, figures that leave a substantial proportion of the observed ethnic differences unaccounted for (Stark, Chernyshenko & Drasgow, 2004).

2.4.4.2 Mixture Latent Variable Modelling

As an alternative to traditional DIF/DTF techniques, a growing number of authors (e.g. Zumbo, 2007; Cohen & Bolt, 2005) have begun to use modern statistical techniques to uncover the nuisance variables that are most strongly associated with differences in responding behaviour, those that might ultimately underlie violations of ethnic measurement invariance.

One such technique that has been recommended for the identification of construct-irrelevant factors that could contribute to ethnic DTF is Mixture Latent Variable Modelling (MLVM; Zumbo, 2007). MLVM (known in the literature, variously, as LVMM, mixture modelling, and finite mixture modelling) is a statistical technique that is designed to identify heterogeneity of response behaviour within a sample. It was originally designed to flexibly model

heterogeneity in a very wide range of random phenomena (McLachlan & Peel, 2000). First developed – as most statistical techniques seem to have been – by Karl Pearson (1894, cited by McLachlan & Peel, 2000) some 120 years ago, it has been successfully applied in fields as diverse as genetics, medicine, engineering, astronomy, biology and marketing (McLachlan & Peel, 2000). It is, at its core, a form of CFA in which each test item is represented by an indicator variable that is made to load on a single latent factor. An example of a mixture model is shown in Figure 8 below:

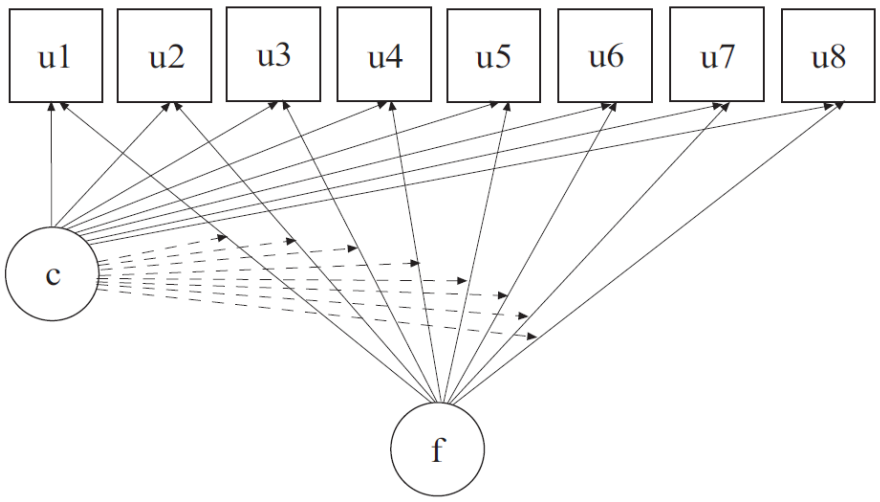

Figure 8. Mixture latent variable model estimated by Mplus for a test consisting of 8 items (from Muthén & Muthén, 2010).

Unlike other CFA approaches, this approach frees all parameters in the model and allows them to vary. Purely on the basis of patterns of their correct and incorrect item responses, participants are categorised as belonging to one of a number of latent classes (represented in Figure 8 as C, a latent variable).

Sawatzky et al. (2012) provide a summary of the mathematics behind MLVM, including the justification by which it may be used for IRT purposes. Mathematically, the probability of a participant responding correctly to a given item, i, according to a mixture model is expressed using a modified version of the equation for the 2PL model. In this equation, the parameter

k represents that factor loadings (λ) and thresholds (τ) are contingent on the latent class to which that participant belongs (represented by C, not to be confused with c_i from the 3PL model):

$$p_{ijk}(Y \geq j|\theta, C = k) = \frac{e^{(-\tau_{ijk}+\lambda_{ik}\theta)}}{1+e^{(-\tau_{ijk}+\lambda_{ik}\theta)}} \quad (4)$$

The equation, therefore, expresses that the probability of a response to an item at or above a particular category, j, (i.e. a correct response) for a given level of θ, depends upon the factor loading and threshold of that item for the latent class to which a participant belongs. When mixture models are applied to an IRT context, these model parameters can be used to estimate conventional IRT parameters. Each item's discrimination parameter (a) can be represented by its factor loading (λ_i) in the model. Difficulty parameters (b) can be calculated by dividing the item's threshold (τ_{ij}) by its value of λ. In other words, the discrimination and difficulty of a given item in a test varies across latent classes of participants, leading to different probabilities of participants of the same level of true ability responding correctly to that item (i.e. that item shows DIF). The item parameters for all classes are estimated simultaneously in the model, and individuals are assigned to a latent class based on the posterior probabilities of each individual belonging to a particular class, the largest posterior probability indicating the most likely class.

Therefore, when MLVM is used to examine ability testing item-level data, each class, theoretically, should have the same level of *g* (represented by θ), but the way in which the test functions for them could give rise to differences in the way in which their level of *g* is measured by the test (represented by variable thresholds and factor loadings across latent classes). Assuming that the correct number of latent classes has been specified in the model, each class should represent a relatively homogeneous pattern of response behaviour within that class.

Zumbo (2007) argues that MLVM procedures are most useful for investigating DTF when used to try to understand the underlying processes that determine response behaviour. By not defining what different latent classes should represent *a priori*, but instead allowing classes to be constructed through the MLVM procedure, key variables can be identified that potentially act as mediators of the relationship between manifest characteristics and test score group differences. Once these key variables have been identified, their prevalence within manifest groups (e.g. different ethnic groups) can be examined to establish whether these key variables can explain the mechanism by which group performance differences in how a test functions arise, without being confounded by within-group heterogeneity.

However, despite the attractiveness of MLVM as a method by which to assess DTF in a test of ability, it has enjoyed only limited success in this capacity in the published literature. Whereas it has infrequently been used to investigate ethnic DTF in the educational psychology literature, it is practically unheard of in the selection literature. In the rare instances where it has been put to use on ability test data, though, relatively little progress has been made in identifying DTF effects large enough to account for ethnic group performance differences.

This is symptomatic of the first of the two key problems in the literature that the present research hopes to address. While much has been published on DIF and DTF as possible explanations of ethnic group test performance differences, these studies have something of a reputation for methodological problems, due in no small part to their relative complexity over most statistical techniques. Hunter and Schmidt (2000) identify three key problems affecting the majority of studies designed to investigate DTF as a plausible explanation of ethnic group test performance differences. Firstly, these studies tend to rely overly on frequentist significance tests. Secondly, they are frequently based on testing data that violates the assumption of unidimensionality, a necessary condition for all DIF/DTF procedures to be carried out. Finally, there is a tendency in studies of this kind to fail to account for measurement error when defining a criterion by which to represent ability.

To properly address this knowledge gap, DTF needs to be examined in a methodologically robust way on modern testing data. This would serve to either rule out DTF as a legitimate explanation for ethnic group differences, or to indicate the extent to which it can explain these differences.

2.4.4.3 The Tension between Measurement Invariance and Prediction Invariance

One final phenomenon that is worthy of attention within the test impact/bias literature concerns the apparent incompatibility between measurement invariance (as defined in section 2.4.5) and predictive invariance. As has been discussed, both conceptualisations of bias represent important considerations for the ethical deployment of testing in selection. However, there is evidence to suggest that – rather than being complementary approaches to studying test invariance across groups of interest – the conditions of measurement invariance (in which a test measures a latent construct consistently across groups) and predictive invariance (in which it predicts some important criterion consistently for those groups) are often mutually exclusive under realistic conditions. Though mathematically complex, Millsap's (1997) argument hinges on the case of two groups that differ in their mean level of a latent trait of interest. If prediction invariance holds for these groups (i.e. the slope and intercept coefficients for both groups are equal), then measurement invariance cannot hold. Similarly, if the test is measurement invariant, regression coefficients must vary across the groups of interest, indicating a lack of predictive invariance. These two forms of invariance, therefore, are fundamentally incompatible.

In spite of the potentially very serious ramifications that this observation has for research in this field, it is interesting to note that, for the most part, it has been paid little attention. Both Borsboom (2006b) and, subsequently, the original author (Millsap, 2007) observe that these findings have been largely ignored, both in the research literature and in practitioner circles.

2.5 Explanations for Patterns of Group Difference

The second key point of disagreement in the literature is the process by which these ethnic group differences arise (and, by extension, what test users and policy makers can do to manage ethnic group test score differences in practice). What might be the underlying causes of ethnic group differences on selection measures has been hotly debated for many years now. Despite this, very little consensus has been reached. In the case of group differences in ability test scores between US Blacks and US Whites, the research world is divided between hereditarian models (e.g. Rushton & Jensen, 2005) and environmental models (e.g. Dickens & Flynn, 2006).

Hereditarian models argue that any group differences in observed ability test scores can be principally accounted for by genetic differences between Black and White test takers. Rushton and Jensen (2005) maintain that, because group differences between Blacks and Whites on cognitive ability measures have not reduced significantly over recent years, and because genetic factors such as mean brain mass and cranial volume parallel IQ so closely, explanations of racial differences that rely solely on differences in socio-environmental factors (which they characterise, somewhat simplistically, as the 'culture-only' position) are inadequate. Additionally, this argument receives some support (albeit indirectly) from the observations – based on twin studies – that the development of verbal ability, spatial ability, perceptual speed and accuracy, and memory appear to depend much more heavily on genetic influences than they do on shared environment (Bouchard et al., 2003).

Conversely, environmental models say that – while genetic differences between ethnic groups exist – social inequalities during development are the principal drivers of the extent to which they are expressed as differences in cognitive ability. Therefore, the interaction between heritable factors and the environment in which one develops account for these differences in the main. Dickens and Flynn (2006) argue that it is a myth that Black-White cognitive ability differences have remained constant over time and that US Blacks have closed the gap between themselves and US Whites on measures of cognitive ability in

recent years. This pattern of narrowing of the difference between Black and White test takers' performances over time seems not to be restricted solely to US sample, evidence for this effect having also been observed in samples from the UK (Woods, Hardy & Guillaume, working paper).

Furthermore, patterns of difference between other social groups have been observed for which the cranial capacity explanation has difficulty accounting. For example, group differences have been found between White females and White males on measures of *g* (Jackson & Rushton, 2006). However, as the difference is only equivalent to a score of around 4 IQ points lower for White females, it can be observed that US White females tend to perform better on ability tests than US Black males do, given that Black males' mean IQ score is around 85 (Rushton, 2000; cited by Rushton & Jensen, 2005). Given that mean brain size is larger for Black US males than it is for White US females, it seems unlikely that group differences in ability test score can be explained adequately by brain size.

More recently, findings from studies into the genetic foundations for cognitive performance identified three distinct single nucleotide polymorphisms (SNPs) that represent common genetic variants associated with common cognitive performance phenotypes (Rietveld, 2014). This would suggest a degree of genetic predisposition to intelligence. However, this particular avenue of research is still very much in its infancy, the authors admitting that, at present, they were only able to distinguish cognitive performance differences amounting to the equivalent of 0.02 S.D. units per each of the three SNPs on a cognitive ability test.

These observations lead to the conclusion that, though an element of the observed Black-White group differences is likely accountable by heritable factors, this is almost certainly much less than proponents of the hereditarian argument would suggest. Instead, these differences may, in fact, be largely due to increases in the level of socio-economic status (SES) and related socio-environmental factors such as access to education and occupational status of Black people in recent years. Consistent with this, there is a

significant support for the hypothesis that socio-environmental factors influence a number of individual differences and important outcomes.

SES is most frequently considered to be a composite construct that is influenced – to a greater or lesser degree, depending on specific conceptualisations – by access to education, access to employment, wealth and prosperity (APA Task Force on Socioeconomic Status, 2007). In examining the potential of socio-economic factors to predict higher education attainment, Sewell and Shah (1967) found a significant positive correlation between academic success and a student's SES based on a weighted aggregate of familial educational level, familial occupation, household annual income and approximate wealth. Developmentally, familial socio-economic status (SES) has found to be strongly positively correlated with ability test score in children (e.g. Noble, Norman & Farah, 2005). Additionally, familial SES during development has been found to exert an influence on how a child's personality develops, those from lower SES background typically developing lower extraversion, openness, emotional stability and conscientiousness (Jonassaint et al., 2011).

To illustrate the impact that socio-environmental differences might have over and above those of heritable factors, it is important to consider how patterns of ethnic difference can be generalised across nationalities. US Whites have been shown to perform significantly better on ability tests than Sub-Saharan Africans. Furthermore, US Blacks also show significantly higher mean ability test scores than Sub-Saharan Africans (Rushton & Jensen, 2005). This would suggest, then, that socio-economic factors such as health and nutrition play a larger role than genetic factors in predicting success on ability tests. Indeed, Jorm et al. (2004) argued that health state and health habits had a mediating effect on gender differences on cognitive ability between men and women in developed world countries, observing that Australian men tended to score higher on certain ability tests and maintained better health habits on average than women. The environmental argument finds further support in the observation, at the national level, that students in Western countries consistently out-perform those from many Middle-Eastern, Latin-American and African nations on measures of *g*, and

that these differences appear to match differences in the relative socio-economic status of citizens of these countries closely (Rindermann, 2007).

However, in spite of the evidence to suggest a strong effect of both present SES and one's socio-economic background during development, no real consensus has been reached on the degree to which ethnic group test performance differences are attributable to either socio-environmental or heritable factors. Due to this lack of agreement in the literature on the causes of ethnic group differences in ability test performance, it is likely that there is something else that can account for these differences, at least in part. To better understand the processes by which these group differences might arise, it is helpful to consider how an individual's test performance can be influenced not only by their level of *g*, but also by factors unrelated to it.

2.5.1 The Relationship between *g* and Test Performance in Classical Test Theory

According to Classical Test Theory (CTT; Novick, 1966), a candidate's observed score (i.e. the score they obtain) on a test can be expressed using the following equation:

$$X = T + E, \quad (6)$$

where X denotes the candidate's observed score, T their true level of ability, and E random error variance.

Reeve, Heggestad & Lievens (2009) extended the CTT model to take into account construct-irrelevant factors that might either help (facilitating factors) or hinder (debilitating factors) a candidate's test performance, thus impacting on their observed score. They formalised this in the following equation:

$$X = T_g + TD + TF + E \quad (7)$$

In this equation, Reeve, Heggestad and Lievens divide T up into three separate sources of *systematic* error variance, along with the single source of random error variance from the previous equation. T_g denotes the variance that can be accounted for by the construct of interest (i.e. *g*), TD the variance contributed by debilitating factors, and TF that contributed by facilitating factors.

The extant literature has made inroads into exploring what some of these sources of systematic variance could be, most frequently in the form of personality traits and the pathways they follow to affect test performance. Chamorro-Premuzic and Furnham (2004) propose a model of how the Big Five personality traits affect test performance in different ways, displayed in Figure 9 below.

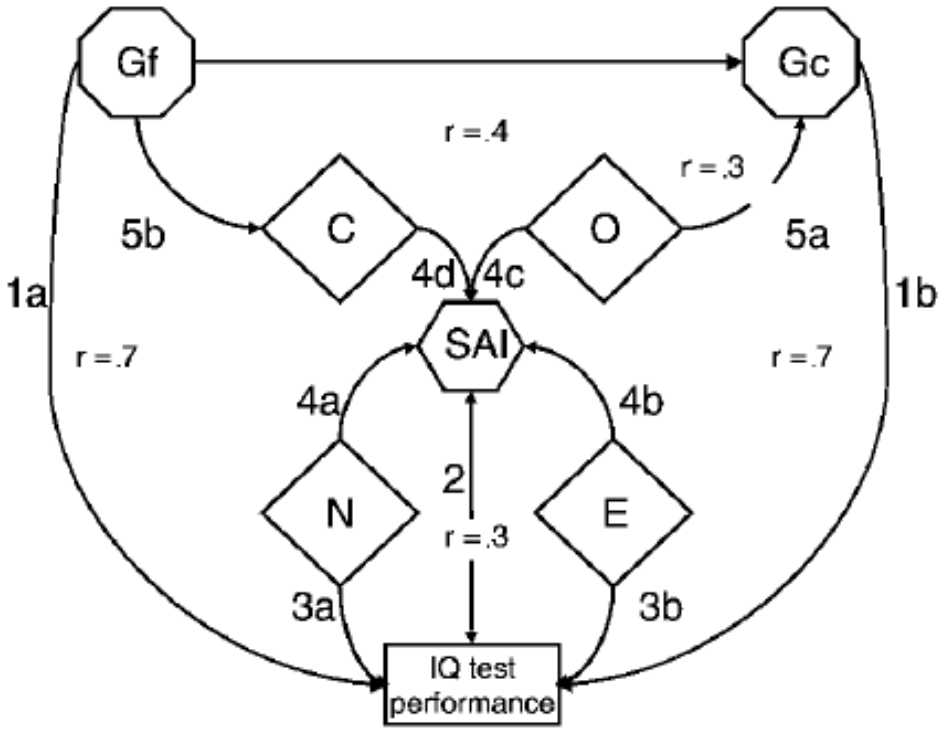

Figure 9. A model for understanding the personality-intelligence interface (from Chamorro-Premuzic & Furnham, 2004).

They argue that whilst some traits affect *g* directly (citing as an example the effect of Openness on the development of g_c), others affect test performance itself while remaining independent of cognitive ability. An example of such a trait is Extraversion, the positive relationship of which with test performance reflects both extraverts' high levels of assertiveness (leading to lower mean item response time, a beneficial factor in the time-pressured environment of timed tests), and low levels of arousal (making them less sensitive to audible distractors, thus increasing concentration) (Chamorro-Premuzic & Furnham, 2004).

A further trait that is known to influence test performance independent of cognitive ability is Neuroticism (versus Emotional Stability). The effect of Emotional Stability on test performance is generally accepted to occur through the mechanism of a candidate's predisposition to experience performance-debilitating test anxiety. The relationship of test anxiety with test performance is well understood thanks to the seminal work of Selye (1956), Lazarus (1966), and those that have built upon their work. It is now accepted that the relationship between anxiety and performance, particularly on complex tasks, follows a Yerkes-Dodson (inverse U-shaped) curve, as depicted in Figure 10 below (e.g. Le Fevre, Kolt & Matheny, 2006).

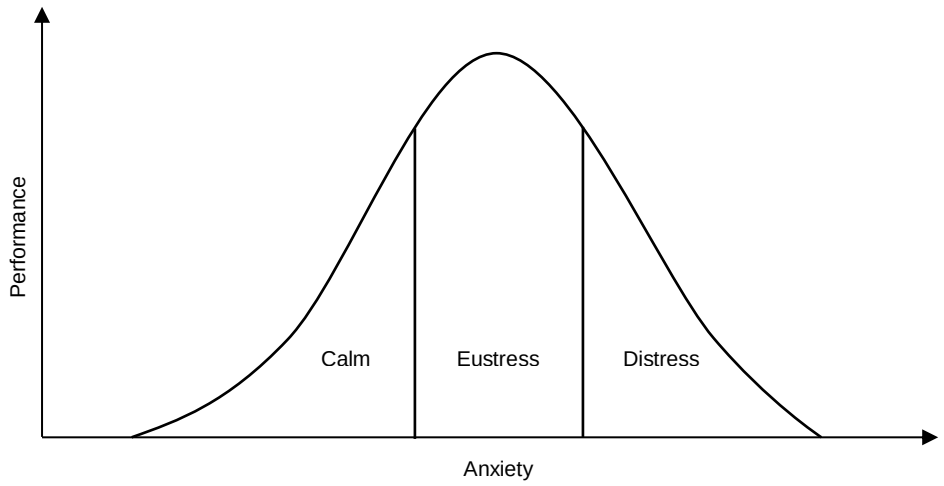

Figure 10. Yerkes-Dodson curve showing the relationship between stress level and task performance.

Moderate stress levels have been demonstrated to actually increase task performance, a phenomenon named *eustress* by Selye (1956). Conversely, high levels of stress have been found to negatively affect task performance (a state called *distress*). An important clarification is made by Selye in that it is not the amount or strength of stressors that cause one to enter a state of distress rather than eustress, rather one's perception and subsequent response to those stressors. Individual differences in emotional stability, therefore, reflect differences in these perceptions of and responses to stressors.

Lupien et al. (2007) offer a biological explanation for why stress levels might impact on cognitive processes. In a review of the literature, they link elevated levels of glucocorticoids (a class of steroid hormone) such as cortisol to cognitive performance, particularly in terms of memory and attention. As these glucocorticoids pass across the blood-brain barrier, they bind to specific receptors, particularly those in the hippocampus, amygdala, and frontal lobes. The binding of glucocorticoids to these receptors appears to affect cognition in a way that mirrors the Yerkes-Dodson curve shown in Figure 10 very closely, enhancing cognitive performance at modest levels, but impeding it in excess. It stands to reason, therefore, that differences in emotional stability have their roots – at least in part – in an individual's sensitivity to these glucocorticoids.

Testing sessions in selection, as with any component of a selection process, are typically high-stakes situations. Candidates are often all too aware that the outcome of the selection process depends upon their performance, and that important aspects of their lives will be affected by this outcome. Testing in selection is naturally, therefore, a highly pressurised environment. Those candidates who are highly emotionally labile (i.e. less emotionally stable), are much more likely to be pushed to the performance-debilitating level of distress than they are to remain in eustress. Conversely, those with a higher level of emotional stability are likely to be more resistant to distress, thus being more likely to remain at the performance-enhancing eustress level.

In the context of ethnic group test performance differences, performance debilitating test anxiety has been most frequently examined in the form of stereotype threat (Steele & Aronson, 1995). Stereotype threat is a phenomenon that has frequently been observed experimentally in which the performance of participants with certain manifest characteristics (such as females and those from ethnic minority groups) may be adversely affected under certain circumstances. According to Steele and Aronson (1995), when a popularly-held stereotype exists that negatively portrays the performance of a social group to which a participant belongs, and group differences between that group and another is made salient in some way, that participant may make self-evaluations of their own ability that could adversely affect their performance. This has most frequently been observed in US Black test takers in educational settings. In these cases (as was the context for Steele & Aronson's initial lab study), if the classically-observed gaps between Black and White student groups in terms of their educational attainment is made salient prior to testing, the impact that this has on their performance becomes something of a self-fulfilling prophecy: Black test takers will perform less well than the White group because they *expect* to do so.

Additionally, there is a degree of evidence that ethnic group performance differences need not necessarily be made explicit for stereotype threat effects to trigger: Nguyen and Ryan (2008) observed that even doing something as seemingly innocuous as recording a participant's ethnicity prior to testing can be enough to create the conditions for stereotype threat to occur.

It is a point of some contention in the literature as to the degree to which stereotype threat can explain ethnic group test score differences. Helms (2005) argues, based on a reinterpretation of Steele & Aronson's original study, that stereotype threat could explain most – if not all – of the 1 S.D. unit difference between Black and White test takers. Conversely, Sacket, Harrison and Cullen (2004) contend that stereotype threat merely exaggerates Black-White test score differences that are already present for whatever reason.

If one were to apply the logic of Chamorro-Premuzic and Furnham's (2004) model to that of Reeve, Heggestad and Lievens (2009), it might give some insight into some of the construct-irrelevant performance facilitating and debilitating factors that could give rise to ethnic group test score differences. These differences could hypothetically be explained in a number of ways. Firstly, the differences in X could be accounted for by ethnic group differences in T_g (i.e. performance differences that were reflective of true differences in intelligence between groups). Alternatively, a candidate's ethnicity – either directly or indirectly – could exert an influence on either TD or TF (or, conceivably, both at once) through the mechanism of group differences in key personality traits.

The existence of meaningful ethnic group differences in personality traits is debated in the literature. Goldberg et al. (1998) observed that US Black, Asian and Hispanic test takers tended to score lower than US Whites in conscientiousness (approximately 0.23 S.D. in all cases), and US Black test takers higher than US Whites on measures of Emotional Stability, though the authors viewed these differences as trivial. In the case of the UK, Ones and Anderson (2002) noted personality differences between British ethnic groups. They observed higher mean scores for extraversion (0.10 S.D.) and lower mean scores for emotional stability (0.28 S.D.) for Asian participants than White participants, lower conscientiousness scores (0.26 S.D.), lower extraversion scores (0.22 S.D.) and higher emotional stability scores (0.11 S.D.) for Black participants than White participants, and lower conscientiousness scores (0.22 S.D.) and lower emotional stability scores (0.10 S.D.) for Chinese participants than for White participants. However, Ones and Anderson admit that these differences are likely too small to represent any meaningful practical effect.

The implication, then, is that even if consistent personality differences exist between ethnicities, the magnitude of these differences is such that the argument that they alone could explain ethnic group test performance differences is difficult to sustain. If it were truly the case that a substantial proportion of the variance in test scores explained by ethnicity could be attributable to construct-irrelevant factors that directly affected test performance,

other traits or characteristics that lay outside personality would also need to be considered. Chamorro-Premuzic and Furnham (2004) allude to one of these in the form of subjectively assessed intelligence (SAI). In their model, SAI represents a person's self-estimate of their own intelligence, which is dependent on a number of factors, both cognitive (i.e. their actual level of intelligence), and non-cognitive (such as confidence in their own ability, which is necessarily linked to Big Five traits such as Extraversion and Neuroticism). As such, SAI is a construct that is related to test performance that is neither truly a personality trait, nor truly a form of intelligence. Chamorro-Premuzic and Furnham note a correlation coefficient between SAI and test performance of .3, implying a moderately strong relationship, yet one that is stronger than that between personality traits and test performance directly. It is possible, however, that this relationship could be better explained by another, related construct that is more objective and less dependent on personality.

2.5.2 Test Familiarity as a Performance Facilitating Factor

Of the non-personality characteristics that could facilitate test performance, the factor that is most frequently alluded to in the literature is test familiarity (Reeve, Heggestad & Lievens, 2009). However, in comparison to test anxiety, the effect of test familiarity on performance is much more poorly researched and understood.

The mechanism by which test familiarity facilitates test performance is most frequently proposed to be by way of an interaction with test anxiety, higher test familiarity helping to reduce test anxiety. However, the foundation of this explanation (and, indeed, of research on test familiarity) is in observations of differences in examination performance in academic settings. The key assumption, here, is that examination performance and ability test performance are influenced by the same antecedents, but this is a difficult argument to support empirically, particularly given the disconnect in the role that Conscientiousness plays in models of both kinds of performance. Whereas Conscientiousness is positively correlated

to academic performance (e.g. O'Connor & Paunonen, 2007; Noftle & Robins, 2007), it is negatively related to test performance (Chamorro-Premuzic & Furnham, 2004).

There is some evidence to suggest that test familiarity and factors related to it have a positive impact on test performance that is separate to the reduction of test anxiety. Some attention has been devoted to investigation of the effect of short-term coaching programmes aimed at improving test performance. While studies of this nature (such as that of Ryer, Schmidt & Schmidt, 1999, cited by Sacket et al., 2001) are rare in the occupational literature, they was a degree of research interest paid to them in the educational literature some time ago. Reviews of the efficacy of coaching programmes has indicated performance increases of approximately 0.1 S.D. units for programmes that focus on the SAT (DerSimonian & Laird, 1983), up to around 0.25 S.D. units for those that focus on tests of achievement. These programmes have even focused on the use of test coaching for the amelioration of ethnic group test performance differences on tests such as the MCAT (Frierson, 1986), though, in a review of the relevant literature, Sackett et al. (2001) conclude that, overall, coaching programmes have demonstrated little or no impact on ethnic group differences. However, almost without exception, these studies have focused on tests that bear little resemblance to the reasoning tests used in selection. Educational admissions tests such as the MCAT and SAT all have substantial components for which performance relies upon knowledge within a subject area. For these tests, training to improve test familiarity will only be of limited benefit. This would, most likely, lead any observed beneficial effects of increasing test familiarity to be underestimated. The implication of this is that the field may well have discounted test familiarity as a cause of meaningful differences in test performance between ethnic groups without having properly investigated it.

To better understand the role of test familiarity as a test performance facilitating factor, it instead needs to be examined in terms of familiarity with ability testing specifically to determine the factors that influence a candidate's familiarity with testing processes. This

could give greater insight into whether ethnic group differences in test familiarity exist, and – if so – what are the root causes of these differences.

An illustration of how test familiarity could directly influence test performance, irrelevant of any effect on test anxiety is illustrated in the following hypothetical example. Consider the following item that could realistically be found on a numerical reasoning test:

> *"Which number comes next in the sequence?*
>
> *0, 1, 1, 2, 3, 5, 8, 13, ___*
>
> *a. 45*
> *b. 21*
> *c. 18*
> *d. 27"*

Although not particularly difficult, participants might reasonably approach this item in different ways, depending on their test familiarity. For example, a participant might look at the progression of numbers in the sequence and, after a while, realise that, for any given number x, the next number is the sum of x and the previous number, in this case making the correct answer b: 21. If he or she did not have experience of testing, the participant might have tested various mathematical (or, even, non-mathematical) hypotheses on the series, eventually arriving at the correct answer. By contrast, if the participant were a mathematician, he or she might have recognised the numbers as belonging to the Fibonacci Sequence, very quickly coming to the realisation that 21 was the next number in the sequence, possibly without performing any mental calculation at all, allowing this participant to very quickly answer the question based on their pre-existing mathematical knowledge. However, a non-mathematician who was familiar with numerical reasoning tests would probably know that this form of basic additive sequence is extremely common in these kinds of tests, so would likely recognise the pattern of the relationship between the numbers more

quickly than the test taker in the first example, allowing this to solve the problem more efficiently than a neophyte test taker.

This pattern recognition, then, is at the core of how test familiarity could affect ability test performance. The mechanism by which this effect could be explained could, therefore, be explained less by its secondary effect on anxiety, but more in terms of Schema Theory (Anderson & Pearson, 1988). There are a theoretically finite number of general forms that test items of any kind can take, so the greater a participant's familiarity with testing, the greater the number of schemata they will hold of what these items are and, consequently, the method that needs to be applied to solving them.

As, according to Schema Theory, schemata are constructed through experience, the root cause of group differences in test familiarity, therefore, must be socio-environmental in nature, in that a person's exposure to ability testing – at least in formal contexts – is likely heavily influenced by the environment in which they develop. Those from higher-socio-economic status (SES) backgrounds are more likely to have attended selective schools, and will likely have more experience in applying and being selected for higher-level jobs, both of which are likely to grant more exposure to ability tests through their use for selection or development. Therefore, at the group level, ethnic differences in test familiarity should be largely accountable by socio-economic status and factors related to it.

There is, to date, no robust, readily-available measure of test familiarity available in the literature. Aside from the tendency for researchers in this area to conflate test familiarity with academic examination familiarity (Reeve, Heggestad & Lievens, 2009), in the rare instances in which ability test familiarity is examined in published studies, it tends not to be assessed by any consistent means. The majority of studies (e.g. Anastasi, 1981; Hausknecht et al., 2007) do not measure test familiarity as a broad construct, instead defining it in terms of practice effects and coaching for specific tests such as the SAT. The tacit assumption of these studies is that the coaching process raises familiarity with that particular test, leading

to performance increases upon retest (Hausknecht et al., 2007). Reeve, Heggestad and Lievens (2009) offer a conceptual definition of test familiarity as captured by these studies:

> *"*[Test familiarity] *(aka Test-specific Knowledge; Test Sophistication) encompasses all construct-irrelevant test-specific "knowledge" that is not associated with the actual ability being measured. Although we recognize there are differences between concepts such as test-taking skills, test-wiseness, test familiarity, and test-specific variance, all of these concepts do share a common theme. Namely, all of these concepts reflect, in various forms, a performance-facilitating factor that is theoretically independent of the ability measured by the test."*
>
> – Reeve, Heggestad & Lievens (2009, p. 34)

However, if test familiarity does, indeed, operate via the mechanism of developing schemata to represent test forms and rules, this definition's focus on test-specificity is too narrow a conceptualisation to properly be investigated by the means normally employed in test familiarity studies. The implication of this is that, if test familiarity is to be examined as a performance facilitating factor, a metric would first need to be developed that defines test familiarity as a broad, persistent construct that generalises across tests, which could be used to measure it in a robust, psychometric way.

2.5.3 A New Model of Ethnic Test Performance Differences

The present research proposes a model of ability test performance based on a synthesis of the models of Chamorro-Premuzic and Furnham (2004) and Reeve, Heggestad and Lievens (2009), tailored specifically to explain ethnic group test performance differences. While it intends to provide an explanation of ethnic group differences in ability test performance, the model is fundamentally socio-economic in nature, in that SES is viewed as a mediator between Ethnic Group and other key predictors of test performance. The proposed model for this study is shown in Figure 11 below:

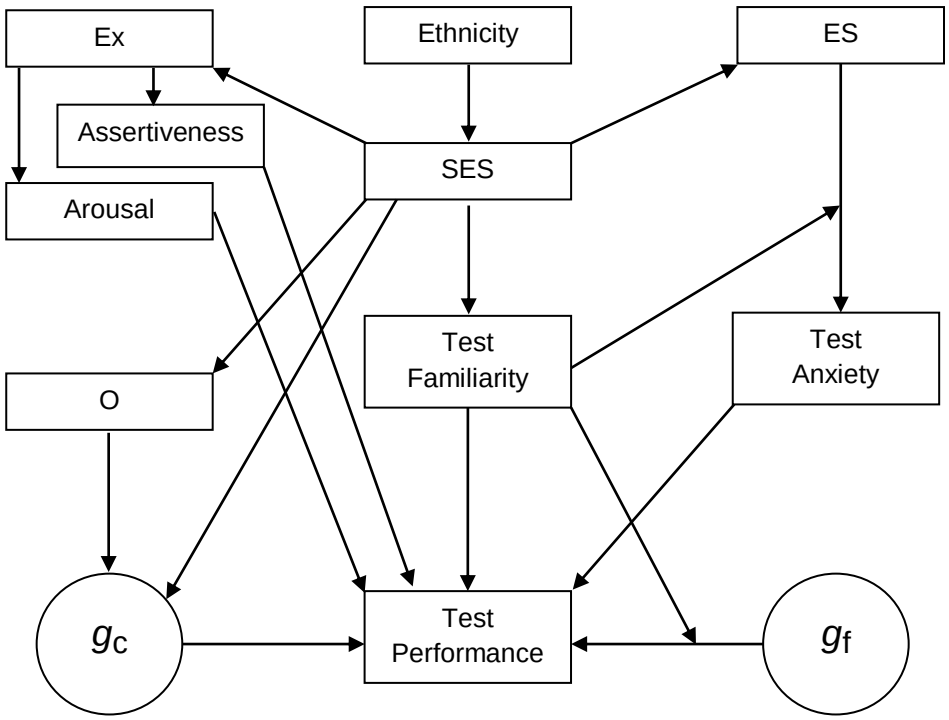

Figure 11. Proposed model showing how facets of *g*, facilitating and debilitating factors interact to bring about ethnic group test performance differences.

As in Chamorro-Premuzic and Furnham's (2004) model, personality traits are related both to the cognitive factors (i.e. facets of *g*), and the non-cognitive ones that are associated with performance on ability tests. However, the totality of evidence does not point to ethnicity in and of itself directly influencing personality variables. Instead, the relationships between ethnicity and the important antecedents of test performance are all mediated by SES. SES is known to be associated much more strongly with differences in these key personality variables (Extraversion, Openness and Emotional Stability; Jonassaint et al., 2011) than is ethnicity itself (e.g. Goldberg et al., 1998; Ones & Anderson, 2002).

In addition to the personality variables associated with test performance from Chamorro-Premuzic & Furnham's model, an additional source of systematic error variance in test performance that is separate from personality is proposed in the form of test familiarity. Test familiarity is proposed to affect test performance in two ways, in both of which it is represented as a construct-irrelevant performance facilitating factor. Firstly, test familiarity is

proposed to reduce a candidate's level of test anxiety (Reeve, Heggestad & Lievens, 2009), so it is included in the model as a moderator of the relationship between a candidate's natural predisposition to react to stressors in a positive or negative way (i.e. their Emotional Stability) and that predisposition to manifest as performance-debilitating test anxiety.

In addition to this, the model also proposes that test familiarity has a direct effect on test performance, the mechanism of which is based on Schema Theory (Anderson & Pearson, 1988). Participants with greater test familiarity will have automatically developed more schemata to represent the general forms that test items can take and – by extension – the strategies necessary to solve them, allowing them to solve the problems presented in test items more efficiently than those unfamiliar with testing. Additionally, these schemata may potentially influence performance by way of pattern recognition. This is represented in the model by test familiarity as a moderator of the relationship between g_f and test performance, in addition to the direct effect that test familiarity is expected to have on test performance.

As the development of test familiarity must come through exposure, it represents a mediator of the relationship between SES and test performance, higher-SES candidates being more likely than lower-SES ones to have encountered testing before in the educational and occupational aspects of their lives, and to have encountered them more frequently.

This model represents a structure by which to address the second key problem facing the field of ethnic bias research. By examining ethnic group test score differences in terms of these performance facilitating and debilitating factors, the degree to which these non-cognitive factors can account for these differences can be established. If it were the case that the greater part of ethnic group differences depended upon differences in these non-cognitive factors, and that – in turn – the root causes of these factors were in socio-economic differences between ethnic groups in society, it would important ramifications for how we understand these issues. Principally, it would serve to bring an end to the

hereditarian/environmental debate, in that ethnic test differences depended far less on heritable factors than the first argument would contend.

2.6 Further Issues in Ethnic Bias Research

Aside from the two key problems facing the field that have been outlined in this chapter, there are a number of issues that consistently cause problems for ethnic bias research, particularly in the context of selection assessment, as is that of the present research. While these issues are not the main focus of the research, they are nevertheless important, as they have important implications for how existing findings in the literature are interpreted, and how future studies should be conducted.

2.6.1 The Spearman-Jensen Effect

A phenomenon that is frequently overlooked in the ethnic bias literature is the Spearman-Jensen Effect (Reeve & Bonaccio, 2009). This effect, first proposed by Jensen (1995), and based on a much older hypotheses outlined by Charles Spearman, focuses on the observation that the magnitude of ethnic group mean test score differences (particularly those between US Black and US White test takers) is directly proportional to the *g*-saturation of a test (that is to say, how highly scores load on to a latent variable representing general mental ability). The obvious solution to this problem would, for practitioners, be to avoid using *g*-loaded tests in selection to ensure that minority ethnic groups are not disadvantaged by their use. However, it was been observed that the predictive validity of a test is directly proportional to its *g*-loading (Reeve & Bonaccio, 2009). This forces practitioners to make compromises between the avoidance of adverse impact (and associated effects) and maximising predictive power.

The Spearman-Jensen Effect is particularly problematic for ethnic bias researchers, as it is typically inferred from this observation that – for whatever reason – White people as a group

typically have higher general mental ability than Black people do. If this were the case, it follows that test scores differences would naturally become more accentuated as *g*-saturation increases, given that higher *g*-loaded tests tap more into the construct of interest than lower *g*-loaded ones do. However, it is possible that the explanation for this phenomenon is more complex than it may appear at first. It may be the case that some aspect of tests that are highly *g*-loaded is responsible for accentuating ethnic group test performance differences. However, to this point, the literature has been unable to provide a plausible explanation for what this might be.

One possible explanation might derive from the assumptions on which it is based. A key assumption of Reeve, Heggestad and Lievens' (2009) modified CTT model is that T_g, TD and TF are separate constructs. As such, the *g*-saturation of a test would only exert an effect on T_g, being entirely independent of TF and TD. However, it is the thesis of the present research that test familiarity impacts upon the way in which *g* is measured (as g_f is a facet of *g*), as it aids pattern recognition in tests, therefore sharing variance with the pattern recognition component of general mental ability. One potential way of understanding how the effect of test familiarity on test performance varies across measures of different types and *g*-saturations is by the differentiation between the variance in test score attributable to g_f and that attributable to g_c, and, by extension, how we conceptualise each. Beauducel, Brocke and Liepmann (2001) have observed that tests that are highly dependent on g_c, such as tests of verbal reasoning, are much more dependent on previous learning processes (including vocabulary, grammatical knowledge and so on). However, there is likely to be an element of pattern recognition to these tests in terms of the general forms of their items, so they will necessarily tap into g_f to some degree. Beauducel, Brocke and Liepmann (2001) argue that the proportion of the variance in test scores for these tests that is attributable to g_f is *irrelevant* to the construct of interest. If this is the case, this would potentially be an

example of how performance facilitating factors such as test familiarity that have classically been viewed as orthogonal to *g* could, in actual fact, interact with it.

If this were the case, theoretically, the following equation would better describe how variance in observed test scores is influenced both by factors that are independent of *g* and of those that are in some way relevant to its measurement:

$$X = T_{g_c} + T_{g_f} + TF_g + TF_i + TD_i + E, \quad (8)$$

where TF_g represents the variance that can be attributed to the component of test familiarity that interacts with the construct of interest yet is irrelevant to its measurement, and TF_i and TD_i represent the variance that can be attributed to those that are truly construct-irrelevant.

In the case of ethnic group differences, this, therefore, provides some possible insight into how differences in test score arise, and offers a possible explanation of the Spearman-Jensen Effect. Due to the potentially shared variance between g_f and test familiarity, it follows that the higher a test's g_f component, the more performance on it will be facilitated by test familiarity.

The theoretical explanation for how mean levels of test familiarity vary between ethnic groups on the basis of differences in exposure due to environment can, therefore, offer an explanation for the Spearman-Jensen Effect. The pattern of ethnic test score differences observed across tests may be accountable less by a test's overall *g*-loading, and more by the degree to which it taps into g_f. The more a test is loaded on g_f, the higher the variance in its scores that is likely to be explainable by test familiarity. The lack of exposure to testing by socially disadvantaged ethnic groups due to the nature of their developmental and subsequent occupational environment would lead, therefore, to their level of g_f to be underestimated. One potential problem with this explanation presents itself in the form of the observation that some tests with very strong g_c components tend to display high *g*-

saturation, such as many verbal reasoning tests do (Schmidt & Hunter, 2004). In these cases, the explanation for the pronounced ethnic group differences on these measures would still be socio-environmental, given the lack of heritability of crystallised ability.

2.6.2 Range Restriction

Range restriction is an extremely common problem associated with bias research in organisational contexts (Cook, 1998). Predictive validity studies tend to compare applicants' test scores at selection with their future job performance. However, by its very nature, selection excludes the applicants with the lowest test scores from the process, all else being equal. This means that future job performance is only – and can only be – measured for those applicants that were successful in the selection process (i.e. the applicants scoring highest on the test). The effect of this is that validity coefficients in studies of this kind tend to be deflated. To combat this, researchers typically estimate the true magnitude of correlation coefficients within predictive validity studies in much the same way as they would for other coefficients with deflated reliability, by applying the Correction for Attenuation (e.g. Salkind, 2010). However, this can only give an approximation of a validity coefficient's true size.

While range restriction is undoubtedly a troublesome issue for much selection assessment research for these reasons, the real problem runs somewhat deeper. When researching ethnic bias in ability testing as a barrier to employment for minority social groups, the accepted method for investigating this is to examine test score data from ethnic groups in organisational settings. However, being job incumbents, participants within these groups would have already been selected for their position. Although often ignored in the literature, this represents a form of range restriction. Any investigation of ethnic bias as a barrier to employment for certain social groups cannot be meaningful without also considering those who have experienced this barrier in real terms: the unemployed. Unfortunately, this section

of the population is frequently overlooked when researching this issue. The upshot of this is that it is likely that many of the negative implications of ethnic bias in selection for the employability of ethnic minority groups are being underestimated.

2.6.3 Representing Ethnicity

A final issue associated with research into ethnic bias – though, arguably, the most fundamental to our understanding of it – is of how ethnicity is defined as a construct. Smith (1986) defined ethnicity as a term used to describe populations who shared a common ancestry and cultural heritage, who are considered by others to be from the same or a similar grouping. This definition has been criticised for being somewhat imprecise and for the observation that the term 'ethnic minority' tends to be applied to groups who are visibly different from the majority group, ignoring those from White minority communities such as Irish and Polish communities (Mason, 2000). In spite of these issues, Smith's definition is a widely used conceptualisation of ethnicity in the research literature (e.g. Kenny & Briner, 2007). However, conceptualising ethnicity in this way misses some of the nuances of this complex construct.

It has been argued that this form of approach to capturing ethnicity ignores the often subtle interactions between the dominant culture of the country in which one lives and that of the society in which one is born and or develops. In a review of the literature, Phinney (1990) noted that in around 2/3^{rds} of studies the concept of ethnicity was not explicitly defined. The remaining 1/3rd could be divided into studies that explicitly defined ethnicity based on one of three broad conceptualisations. These were approaches based on social identity theory, those based on identity formation and those based on the concepts of acculturation and culture conflict. The last of these conceptualisations argues that ethnic identity is only meaningful in environments in which two or more ethnic groups are in contact over a period of time. This approach categorises individuals from an ethnic minority group who are

coexisting with others from a different culture to theirs as belonging to one of four groups based on how strongly they identify with both their own culture (defined as the 'ethnic group') and the culture of the society in which they live (defined as the 'majority group'). If an individual identifies strongly with both the majority group and the ethnic group, they are said to be acculturated, integrated and bicultural. If they identify strongly with the majority group but weakly with the ethnic group, they are said to be assimilated. If they identify weakly with the majority group but strongly with the ethnic group, they are said to be ethnic identified, ethnically embedded, separated and dissociated. Finally, if they identify weakly with both the majority group and the ethnic group, they are described as being marginal.

For example, a participant from Eastern Europe may have lived in the UK for many years, but depending on the strength of their identification with both UK culture and Eastern European culture, they might view themselves as assimilated within UK culture, dissociated from it, or bicultural. The implication of this is that simply to label this person as 'White' would be an oversimplification, but, equally, to label them as 'Non-British White' might also fail to capture their ethnicity accurately, particularly if, over time, they had become highly assimilated within the culture of the UK. This conceptualisation of ethnicity is preferable to those most frequently used (e.g. Smith, 1986) as it recognises that ethnic identity can be defined by both the cultures of the majority group and the minority group simultaneously, rather than being simply a facet of a person's cultural heritage. However, it is a rarity in the ethnic bias literature for this distinction to be made.

CHAPTER 3: STUDY 1

3.1 Study Overview

The first study aimed to establish the extent to which ethnic group test performance differences could be accounted for by Differential Test Functioning (DTF). To address this aim, it was judged to be most appropriate to focus on tests that are well known in the literature to be highly *g*-loaded and to exhibit ethnic group test score differences. Raven's Progressive Matrices are a family of abstract reasoning tests that have been observed to be highly *g*-loaded (Ashton & Lee, 2006), and that also tend to display large differences in mean test score between White and Black test takers. There is evidence to suggest that this difference is typically around 1 S.D. unit in favour of the White group in US samples (e.g. Jensen, 1998), though Owen (1992) observed a difference of almost 3 S.D. units between White and Black South African secondary school pupils. Given the tendency of the literature to focus on Black-White differences in the US, it was reasoned that it would be more fruitful to compare a wider range of ethnic groups across national groups within a global sample, rather than focus purely on a White-Black dichotomy. This would allow greater understanding of the nature of ethnic group differences that would ultimately be much more generalisable than much of the research on ethnic differences in the literature.

For this study, two data archives were obtained from Pearson Talent Lens, the UK distributor of the Raven's series. The first was an archive of item score and test score data for the Raven's Standard Progressive Matrices (SPM; N = 1880). The second was a similar archive for Raven's Advanced Progressive Matrices (APM; N = 1601). Both datasets additionally held data on participants' gender, ethnic origin, country of birth, occupational group, highest educational qualification, and religion (though the final of these was deleted before analysis as it was judged to be irrelevant for the purposes of the present study).

The SPM archive was analysed first. The aim of the analyses was, firstly, to establish whether evidence existed of ethnic group test score differences. Having observed ethnic

group differences, two competing methods of DTF identification were used to investigate whether these differences could be accounted for by DTF and – if they could – to what extent these effects could be explained by secondary variables, particularly those related to socio-economic status. The APM archive was then analysed using the same strategy in an attempt to replicate the findings of the analyses of the first archive.

3.1.1 Hypotheses

On the basis of the broad aim of the study, the following hypotheses were generated:

H_1: Differences in group mean total test score will exist between ethnic groups.

The overwhelming majority of research indicates that ethnic group mean test score differences exist (e.g. Cooper & Robertson, 1995), and that these differences, while gradually decreasing over time, are still present today (e.g. Dickens & Flynn, 2006). There is no plausible reason why this should not be expected to be the case in the present study, at least to some degree.

H_2: The LR Method will not be sensitive enough to identify ethnic DTF in the samples.

H_3: Using a MLVM approach, ethnicity will be significantly related to heterogeneity in response behaviour.

The flawed assumption of traditional ethnic DIF research is of within-group homogeneity of response behaviour. Bound up in this assumption is the implication that it is something to do with ethnicity itself that is responsible for making someone approach a test in a particular way, something that is present for all members of particular groups, but absent for all members of others. However, the failure of traditional ethnic DIF techniques to adequately

explain test score differences suggests that this is not the case. It is, therefore, predicted that the LR Method will not yield compelling evidence of ethnic DTF. However, a degree of difference between ethnic groups in their response behaviour must exist for previous claims of ethnic DTF to be plausible. It is, therefore, predicted that ethnic group will be weak to moderately – yet significantly – related to the response group classification generated in MLVM.

H_4: Socio-environmental factors will be able to account for any differential test functioning detected.

As MLVM controls for true ability, it is expected that socio-environmental factors will be most responsible for differences in responding behaviour. Furthermore, ethnic groups will display group differences on measures of these factors, suggesting that SES mediates the relationship between ethnicity and test score group differences.

3.2 Raven's SPM Data Archive

3.2.1 Measures

Raven's Standard Progressive Matrices

Raven's Standard Progressive Matrices (SPM) is a well-validated and widely-used measure of abstract reasoning. First published in 1938, it is distributed in the UK by Talent Lens, the occupational testing arm of Pearson Assessment.

The SPM consists of 3 practice items, followed by 28 test items, all of which follow a common format. For each item, test takers are presented with a 3x3 matrix of abstract shapes, all of which are related to one another is some way. The bottom-rightmost shape in each matrix is missing. In each item, the task for participants is to select the shape that

completes the matrix from a choice of eight shapes. An example of what one might expect to see within an item on the SPM is shown in Figure 12:

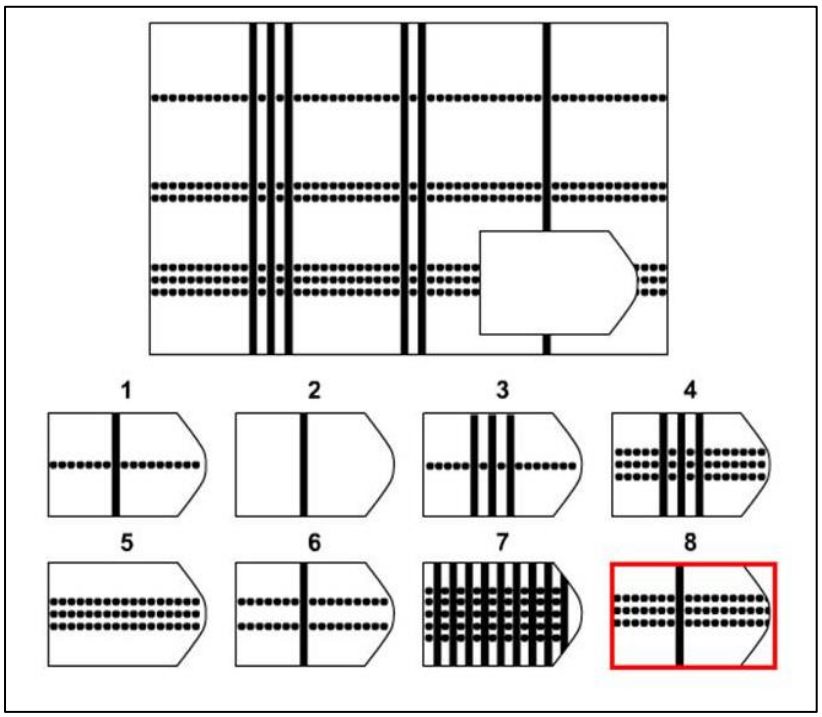

Figure 12. Example item from the Raven's SPM (used with kind permission from Pearson Education Ltd.)

For the item in Figure 12, a participant must choose the shape that completes the matrix from eight options. Examining the matrix, the pattern of solid, vertical lines decreases from 3 to 1 as the columns in the matrix move from left to right. At the same time, the pattern of horizontal lines increases from 1 to 3 as the rows in the matrix move from top to bottom. Looking at the possible response options, option 8 is the only option that could logically complete both the horizontal and vertical patterns within the matrix. This item is, of course, very straightforward to solve (and is, in fact, used as a practice item before the test begins), though the items rapidly become much more complex, requiring participants to employ (in theory) increasingly higher levels of abstract reasoning to deduce the correct shape.

Though the SPM is timed, the time allowed is 47 minutes, meaning that, in practice, most candidates complete all the questions before the time limit.

The Raven's measures have demonstrated excellent levels of construct validity, displaying correlation coefficients with the WAIS matrix reasoning scale scores and overall scores between .74 and .84 (Talent Lens, 2009). Evidence shows that the Raven's tests measure g_f almost exclusively, having very little g_c component (Carpenter, Just & Shell, 1990). However, it has been suggested that success in solving some of the more complex problems within these measures is influenced by working memory capacity, albeit to a lesser degree (Unsworth & Engle, 2005). The measures have demonstrated good levels of discriminant validity with other measures, such as measures of spatial ability (Schweizer et al., 2007).

According to its literature (Talent Lens, 2009), the Raven's series is useful in organisational contexts to measure the capacity of candidates for analysing and solving problems using complex novel information, making it useful when predicting future job performance on a variety of job tasks. It also offers a measure of a candidate's ability to learn, meaning it can provide insight into how easily they might absorb future job-relevant training. For these reasons, both the SPM and the APM have been very widely used in organisational settings, both in the UK and globally. It is most commonly used to assess candidates in supervisory and entry–level management roles, and for those in mid-level consultancy roles (Talent Lens, 2009).

Within the dataset, each participant's total raw score on the SPM was recorded, along with the number of items attempted and the percentile score to which their score equates against the UK population norm group used to score the tool. For the purposes of IRT analyses, dichotomously scored item response data (i.e. 0 = 'incorrect' and 1 = 'correct') was recorded for each participant's response to the individual test questions. As IRT methods typically do not cope well with missing data, if a participant did not attempt a test item, that item was scored as incorrect.

Demographic Variables

In addition to the test data, a number of demographic variables were held in the data archive. Participant gender was recorded dichotomously (as 'male' or female'). Ethnicity was measured by asking participants to indicate which group they felt they identified as belonging to from a list that closely matched the standard question recommended by the Office for National Statistics to assess ethnicity (ONS; 2011). Nationality was measured as a free-response variable in which participants were asked to enter their country of birth (though they did not have to indicate their country of current residence or, indeed, how long they had remained in that country after their birth). Educational level was recorded by asking participants to select their highest level of occupational qualification, which was then represented within the dataset as a rating on a 7-point ordinal scale from 1 ('No formal qualifications') to 7 ('Doctoral Degree'). Additionally, participants' occupational group and tenure within their current role was recorded.

3.2.2 Sample

Data in the archive was held for 1880 participants. Nine hundred and eighty-eight participants described their gender as male (52.6%), 430 as female (22.9%), and 462 did not report their gender (24.6%). Two hundred and fifty-one participants reported their ethnicity as 'White – British' (13.4%), 22 as 'White – Irish' (1.2%), 168 as 'Any other White background' (8.9%), 17 as 'Black or Black British – African' (0.9%), 6 as 'Black or Black British – Caribbean' (0.3%), 131 as 'Asian or Asian British – Indian' (7.0%), 47 as 'Asian or Asian British – Pakistani' (2.5%), 13 as 'Asian or Asian British – Chinese' (0.7%), 8 as 'Asian or Asian British – Bangladeshi' (0.4%), 70 as 'Asian or Asian British – Any other Asian background' (3.7%), 172 as 'Middle Eastern' (9.1%), 1 as 'Latin-American' (0.1%), 80 as 'Mixed – White and Asian' (4.3%), 6 as 'Mixed – White and Black Caribbean' (0.3%), 22 as 'Mixed – White and Black African' (1.2%), 80 as 'Mixed – White and Asian' (4.3%), 23 as 'Any other mixed background' (1.2%), 15 as 'Other ethnic group (not specified)' (0.8%), and

827 (44.0%) did not report their ethnicity. Within the dataset, there were 63 countries of birth reported by the participants. Relatively few of these countries represented a substantial proportion of the total sample, many numbering fewer than 10 participants. The most frequently reported countries of birth were (in descending order) the UK (273; 14.5%), Oman (262; 13.9%), India (128; 6.8%), Qatar (83; 4.4%), Portugal (79; 4.2%), and Saudi Arabia (65; 3.5%). Four hundred and seventy-nine participants (25.5%) did not report their country of birth. Participants represented a wide range of occupational backgrounds, the most commonly reported of which were engineering (87; 4.6%), students (105; 5.6%), HR (62; 3.3%), marketing (59; 3.1%), and customer service (55; 2.9%).

3.2.2.1 Conceptualising Ethnicity

Due to the large number of ethnic categories within the dataset, coupled with the low frequencies within many of these categories, it was necessary to reconceptualise the way ethnicity was represented in the dataset so that meaningful comparisons could be made between groups. These groups needed to be large enough to allow for reasonable statistical power in the planned analyses. However, as discussed in section 2.6.3, a definitive set of rules by which to categorise participants' ethnicity is far from agreed upon in the literature.

The present study was limited by the background data available in the archive, meaning that cultural differences – particularly those described by Phinney (1990) in terms of a person's perceived identification with the culture in which they live if different from their culture of origin – were impossible to deduce. As a result, the present study was forced to take a pragmatic approach to ethnic group classification, one that simplified ethnicity to groups that would be made up by enough participants to allow for robust statistical analyses to be conducted. While the author recognises that this approach misses some of the nuances of ethnicity as a construct, it was deemed necessary for the purposes of the study.

A particular difficulty arises with how to manage data from participants of mixed ethnicity. A current political trend in the UK is to refer to people from many minority ethnic groups as 'BAME' (an acronym that stands for 'Black And Mixed Ethnic'). While this term is viewed by some as a modern and politically correct shorthand to refer to the breadth of minority ethnic groups in the UK (as opposed to the White majority), its use as a method of categorising people is reductionist in the extreme. Within 'mixed ethnicity', there are a huge number of variations in an individual's ethnic heritage (e.g. White and Black, White and Asian, White and Middle Eastern, Black and Asian, Black and Middle Eastern, and so on), the degree to which one ethnic group is represented within that person's direct ancestry versus the other (or others), the cultural influence of their household and community, and the degree to which that individual identifies with each ethnic and/or cultural group of which they could be deemed a member. All of these factors are likely to produce extreme heterogeneity within even a single mixed ethnic classification, making any sensible interpretation of patterns within data for these groups to be next to impossible.

On the basis of these issues, the following strategy was used to categorise participants' ethnicity. At the broadest level, ethnicity was recoded into a dichotomous variable, representing participants as either White (White – British, White – Irish and any other White Background) or Non-White (all other participants who had chosen to report their ethnicity). Beyond this, for the sake of simplicity, participants of mixed ethnicity were excluded from ethnic classification, the confounding effects of the 'mixed ethnic' construct being deemed to be too problematic to reasonably account for. Participants were additionally classified into a polytomous nominal variable, representing the broad ethnic classifications of White, Asian, Black, Middle Eastern and Latin-American (though the last of these categories was excluded from analyses on the basis of having a sample of just a single participant). Finally, a further reclassification of ethnicity was made in an attempt to find some reasonable middle ground between the broad-stroke classification of the previous two variables and the large number of ethnic groups recorded in the dataset. In this variable – hitherto referred to as 'reduced

ethnic group classification' – ethnic categories were retained from the original set of classifications if they fulfilled two criteria: To be retained, an ethnic category had to both represent a meaningful proportion of the total sample who had reported their ethnicity (> 2.5%), and have a distribution of test scores that approximately matched that of the total sample in terms of its dispersion (i.e. similar standard deviation, range, skewness and kurtosis). The criterion of 2.5% is, ultimately, an arbitrary one based on the size of the sample and the number of ethnic categories within that sample, but was chosen to ensure a balance between inclusion of ethnic groups and representativeness of those groups selected. Of these ethnic categories, 6 fulfilled both of these criteria: White – British, White – Any other White background, Middle Eastern, Asian or Asian British – Indian, Asian or Asian British – Pakistani, and Asian or Asian British – Any other Asian background.

3.2.3 Results

Unless otherwise stated, all analyses were conducted in IBM SPSS Version 20. On the recommendation of Zumbo (1999), prior to analysis, all cases that included missing data for ethnicity were removed from the dataset, as missing data is problematic for procedures that rely on logistic regression. This left a total of 1053 cases in the cleansed dataset.

3.2.3.1 Ethnic Group Test Score Differences

To address the hypotheses for this study, the first step was to establish if meaningful ethnic group test score differences existed in the dataset. If these could be established, ethnic DTF could then be examined as a potential explanation for these differences. Mean group test scores were calculated for each of the three conceptualisations of ethnicity. Additionally, Cohen's *d*, a measure of the mean difference in pooled standard deviation units between two groups is shown for each minority group relative to the White majority, as well as the percentile to which each subgroup's mean raw score equates (when rounded to the nearest

whole number). These are shown in the following tables. As there was only a single participant in the Latin-American broad ethnic group, this category was excluded from this and all subsequent analyses.

Ethnic Group	N	Mean SPM Raw Score (S.D.)	Percentile	Cohen's *d* relative to White group
White	441	22.11 (4.59)	57th	-
Non-white	612	20.65 (5.01)	48th	0.30

Table 1. Mean raw score, percentiles and Cohen's *d* by ethnicity classified as White/Non-white.

Ethnic Group	N	Mean SPM Raw Score (S.D.)	Percentile	Cohen's *d* relative to White group
White	441	22.11 (4.59)	57th	-
Asian	269	21.04 (4.87)	48th	0.27
Black	24	17.17 (6.27)	25th	0.90
Middle Eastern	172	21.55 (4.58)	57th	0.12

Table 2. Mean raw score, percentiles and Cohen's *d* by broad ethnic group classification.

Ethnic Group	N	Mean SPM Raw Score (S.D.)	Percentile	Cohen's *d* relative to White – British group
White – British	251	21.50 (4.97)	57th	-
Other White	168	23.04 (3.88)	67th	-0.35
Middle Eastern	172	21.55 (4.58)	57th	-0.01
Indian	131	21.27 (4.65)	48th	0.05
Pakistani	47	20.98 (5.90)	48th	0.10
Other Asian	70	21.14 (3.81)	48th	0.08

Table 3. Mean raw score, percentiles and Cohen's *d* by reduced ethnic group classification.

The data in Tables 1, 2 and 3 indicate that there are consistent test score group differences between White test takers and those from minority ethnic groups. In Table 1, the mean difference in performance between White and Non-white test takers is around 1.5 raw score points. At first glance, this does not appear to be a huge difference, though it equates to a difference in mean score of around 0.3 S.D. units. Furthermore, this difference has some practical significance, in that it signifies a difference in percentile score between the White majority at the 57th percentile and Non-white test takers at the 48th percentile. Though this is

not enough to be judged as having adverse impact on the Non-white group, it would still potentially lead to differing selection ratios between these groups in a selection process with a cut score at the 50th percentile.

This pattern is also present in Table 2, the White majority group consistently out-performing the minority ethnic groups. The practical effect sizes of these differences are relatively small between Asian and Middle Eastern test takers and the White majority, though it might again equate to differences in selection ratios in the circumstances of some selection processes. These differences are most pronounced for the White majority and Black test takers. Though the sample size of Black test takers is relatively small, the dispersion of its raw score is relatively similar to that of the dataset as a whole, showing similar skewness, kurtosis, range and only a slightly larger standard deviation. The difference between these groups of five raw score points represents a difference of slightly less than 1 S.D. unit. Furthermore, this difference is equivalent to a percentile difference between the 57th percentile for the majority and the 25th percentile for the Black group. This might potentially represent adverse impact on the Black minority group, were it to be deployed in a selection process.

Table 3 offers further insight into ethnic group differences, the most striking of which is between the White British group and the group of test takers of any other White background (excluding White Irish test takers). While there are relatively minor differences between the other ethnic groups and the White British majority, the non-British White group appear to score substantially better than the majority group, a difference of around 0.35 S.D. units. The implication of this is that the presence of the White British majority might, in actual fact, be suppressing differences between White test takers and those of minority ethnic groups.

To investigate these differences further, a series of simple linear regressions was conducted to establish to what extent ethnicity could explain variance in total test scores. The two polytomous ethnicity variables were first recoded into dummy variables to represent specific ethnic groups. The results of these analyses are shown in the following tables:

	B	S.E. B	β
Constant	20.65	0.20	
Ethnicity	1.46	0.30	0.15*

Table 4. Linear regression of ethnicity reclassified as White-Non-white predicting total SPM raw score. Note: R^2 = .02; *p < .001

	B	S.E. B	β
Constant	26.00	4.72	
White	-3.89	4.73	-0.41
Asian	-4.96	4.73	-0.47
Black	-8.83	4.82	-0.30
Middle Eastern	-4.45	4.73	-0.36

Table 5. Linear regression of broad ethnic group predicting total SPM raw score. Note: R^2 = .03

	B	S.E. B	β
Constant	21.50	0.30	
Other White	1.53	0.46	0.13*
Indian	-0.24	0.50	-0.02
Pakistani	-0.52	0.73	-0.03
Other Asian	-0.36	0.62	-0.02
Middle Eastern	0.05	0.46	0.004

Table 6. Linear regression of reduced ethnic group classification predicting total SPM raw score.

Note: R^2 = .02; *p < .001

Examining Tables 4, 5 and 6, all conceptualisations of ethnicity can explain small yet significant variance in total SPM raw score. Ethnicity classified as White-Non-white can explain around 3% of the variance in raw scores, White test takers scoring 1.47 raw score points above Non-white test takers in the model. Broad ethnic group classification explained around 2% of the variance in raw scores. Though none of the betas for individual ethnic groups were significant, the model was significant overall. Similarly, reduced ethnic group classification explained 2% of the variance in raw scores. Most betas were not significant, though the beta for the White – Any other White background group was significant at the 99.9% level. In this model, the White – British term was excluded from the model due to its collinearity with the Middle Eastern term (presumably due to these groups' mean scores and distributions being so similar). However, the overall model was significant. Overall, these findings support H_1.

3.2.3.2 Traditional Ethnic DTF Identification using the LR Method

Having discovered evidence of meaningful group differences between ethnic groups, the analyses continued by exploring whether evidence could be found of DTF. Two methods were employed to do this. The first is a traditional method of DIF detection (which can subsequently be used to detect DTF based on the contribution of each individual item's DIF effect) known as the Logistic Regression (LR) Method.

The LR Method (e.g. Zumbo, 1999) is perhaps not as well-known as some DIF detection methods such as the Mantel-Haenszel Procedure or the SIBTEST, but it does have some clear advantages for bias researchers over these methods (e.g. Fidalgo, Mellenbergh & Muñiz, 2000). French and Maller (2007) have demonstrated that, while it tends to produce similar results to these more popular methods, it is more sensitive to the detection of non-uniform DIF. An additional advantage to the LR Method over its competitors is that it is relatively straightforward and can be conducted without specialist statistical software (though it can be quite time consuming and calculation-heavy for the researcher, very little of the process being automated).

The method tests for DIF in each item by regressing item score (a dichotomous categorical variable representing correct or incorrect item response) on to ability (represented in the first instance by total test score), group membership (for example, ethnic group, investigating ethnic DIF) and the interaction term between these two variables. The Chi Squared statistic for the full model is then compared to those when the group and interaction terms are removed. DIF is indicated by significance of the difference in the value of Chi Squared statistics for the full model (*df* = 3) and the model with only the ability term included (*df* = 1). If this value of Chi Squared for this part of the model (i.e. *df* = 2) is larger than the upper-tail critical value for significance for 2 degrees of freedom, DIF is detected. The type of DIF is indicated by the nature of the change in R^2 over and above that explained by the ability term as the group and interaction terms are added sequentially to the model. If the majority of the change in R^2 occurs when the group term is added to the model, it is indicative of uniform

DIF. Conversely, if the majority of the change in R^2 occurs when the interaction term is subsequently added, the DIF is non-uniform in nature (French & Maller, 2007).

In addition to the use of the Chi Squared statistic as an indicator of DIF, it is recommended that attention is paid to the magnitude of change in R^2. Even if the change in Chi Squared is significant, a small change in R^2 would indicate that the magnitude of the DIF effect observed is of little practical significance (Zumbo, 1999). Consideration of the effect size of DIF (named the Zumbo-Thomas effect size) has been demonstrated to lead to fewer Type I errors than when the Chi Squared test alone is used to detect DIF using this method (Jodoin & Gierl, 2001). Though there are a number of critical values for this effect size recommended by researchers, French and Maller (2007) observe that those recommended by Jodoin and Gierl (2001) have shown greater accuracy in detecting DIF over the minimum R^2 of .130 originally recommended by Zumbo and Thomas (1997). Jodoin and Gierl (2001) recommend that a change in R^2 of less than .035 indicates negligible DIF, a change in R^2 between .035 and .070 a moderate DIF effect, and a change in R^2 of greater than .070 a large DIF effect.

As with many traditional DIF detection approaches, the LR Method uses a purified criterion of true ability to increase the power of the DIF analysis (French & Maller, 2007). This purified criterion is typically one in which items that display ethnic DIF are iteratively removed from the criterion until the resultant test score only consists of items that display no ethnic DIF. After the initial iteration of DIF tests, the analysis is then run again for all items using a modified measure of ability, calculated by subtracting score on the items previously flagged for DIF (n) from total test score (N). To reduce the likelihood of Type I errors in this second iteration (and all subsequent ones), Holland and Thayer (1988) additionally recommend that the item score for the item being analysed be included in the refined ability term when testing that particular item, even if it has been flagged for exclusion in the previous iteration. This process is repeated until the same set of DIF items are detected in two consecutive analysis iterations or no other items are identified (French & Maller, 2007). The final value of N - n in

this final iteration represents the purified criterion of ability to be used in the main analysis. The main analysis is then run in the same manner as for the purification process, using the newly purification criterion.

The LR Method was used to detect ethnic DIF in the SPM dataset. As traditional DIF requires a dichotomous classification of ethnicity (to represent the dichotomy between focal and referent ethnic groups), the White-Non-white variable was used as the group variable for the analysis. The interaction term between total raw score and White-Non-white group was computed. A series of logistic regressions was then run. For each item in the SPM, the item score was the dependent variable. In Step 1, total raw score was entered into the model. In Step 2, the group variable was entered into the model. Finally, in Step 3, the interaction term was entered into the model.

The output of each logistic regression model was then examined to determine whether or not ethnic DIF could be identified for that item. To be flagged as having DIF, the analysis had to meet two criteria. Firstly, an item was required to have a *p* value of less than .01. A robust *p* value was chosen in this case to mitigate the family-wise error rate associated with running large numbers of regression models, thereby lessening the likelihood of Type I errors. This is particularly important in LR method DIF detection, as there is evidence to suggest that the purification process can inflate the likelihood of Type I errors (French & Maller, 2007). Secondly, the analysis would need to show a non-negligible effect size using Jodoin & Gierl's (2001) criteria for judging the magnitude of Nagelkerke's R^2 (a measure of pseudovariance used in logistic regression that is analogous to the traditional R^2 measure of explained variance used in linear regression). If both of these criteria were met, the item was judged to display DIF.

SPM Iteration 1

	Nagelkerke R^2 values at each step in the hierarchical regression					
	Step 1	Step 2 (uniform DIF)	Step 3 (non-uniform DIF)	DIF $\chi^2(2)$ test	DIF R^2 (ΔR^2 between Steps 1 & 3)	DIF detected
Item 1	.335	.355	.355	10.38**	.020	Negligible
Item 2	.169	.175	.176	3.42	.007	Negligible
Item 3	.163	.165	.165	0.59	.002	Negligible
Item 4	.246	.246	.246	0.08	.000	Negligible
Item 5	.250	.250	.250	0.31	.000	Negligible
Item 6	.221	.221	.223	1.14	.002	Negligible
Item 7	.383	.384	.386	2.86	.003	Negligible
Item 8	.378	.379	.379	1.23	.001	Negligible
Item 9	.330	.334	.334	4.23	.004	Negligible
Item 10	.475	.479	.479	5.57	.004	Negligible
Item 11	.294	.295	.295	0.83	.001	Negligible
Item 12	.275	.276	.276	0.47	.001	Negligible
Item 13	.338	.344	.346	4.63	.008	Negligible
Item 14	.275	.276	.277	1.84	.002	Negligible
Item 15	.241	.242	.246	4.66	.005	Negligible
Item 16	.329	.331	.332	3.66	.003	Negligible
Item 17	.332	.333	.333	0.72	.001	Negligible
Item 18	.528	.528	.529	0.35	.001	Negligible
Item 19	.477	.478	.478	0.47	.001	Negligible
Item 20	.509	.511	.511	1.83	.002	Negligible
Item 21	.507	.508	.508	0.68	.001	Negligible
Item 22	.455	.455	.455	0.13	.000	Negligible
Item 23	.292	.293	.294	1.73	.002	Negligible
Item 24	.442	.442	.445	3.80	.003	Negligible
Item 25	.453	.454	.454	1.17	.001	Negligible
Item 26	.451	.453	.462	12.00**	.011	Negligible
Item 27	.441	.445	.453	11.89**	.012	Negligible
Item 28	.311	.311	.331	18.18***	.020	Negligible

Table 7. Initial iteration for LR Method identification of White-Non-white ethnic DIF. Note: *p < .05; **p < .01; ***p < .001

Though significant ethnic DIF was detected for some items in the first iteration, these represented negligible practical effects. The analysis, therefore, terminated. This might lead one to assume that the SPM did not display any DTF for White and Non-white groups. However, just as many DIF effects may cancel one another out at the test level (Gibson & Harvey, 2003), it is entirely possible that many negligible DIF effects could combine to produce an overall non-negligible DTF effect. For this reason, DTF was calculated based on the parameters generated by the LR analyses, following the method described by Oliveri, Olson, Ercikan and Zumbo (2012). In this approach to DTF identification, a series of DIF

plots are created for each item, based on the regression coefficients obtained in the previous analyses. Predicted item responses are generated for each group based on every possible total test score and plotted to generate a set of item characteristic curves (ICCs) for both focal and referent groups. These curves are then added together to generate test characteristic curves (TCCs) for both groups, the x-axis of which represents observed total score, and y-axis predicted total score. The advantages of this method are, firstly, that it generates a clear visual representation of the DTF effect at every level of test performance, and, secondly, that it allows for easy interpretation of these effects, in that the real impact of DTF at a particular observed total score is simply the difference in predicted total score between groups, measured on the graph in raw score points.

Figure 13 below shows the resultant TCCs for White and Non-white groups:

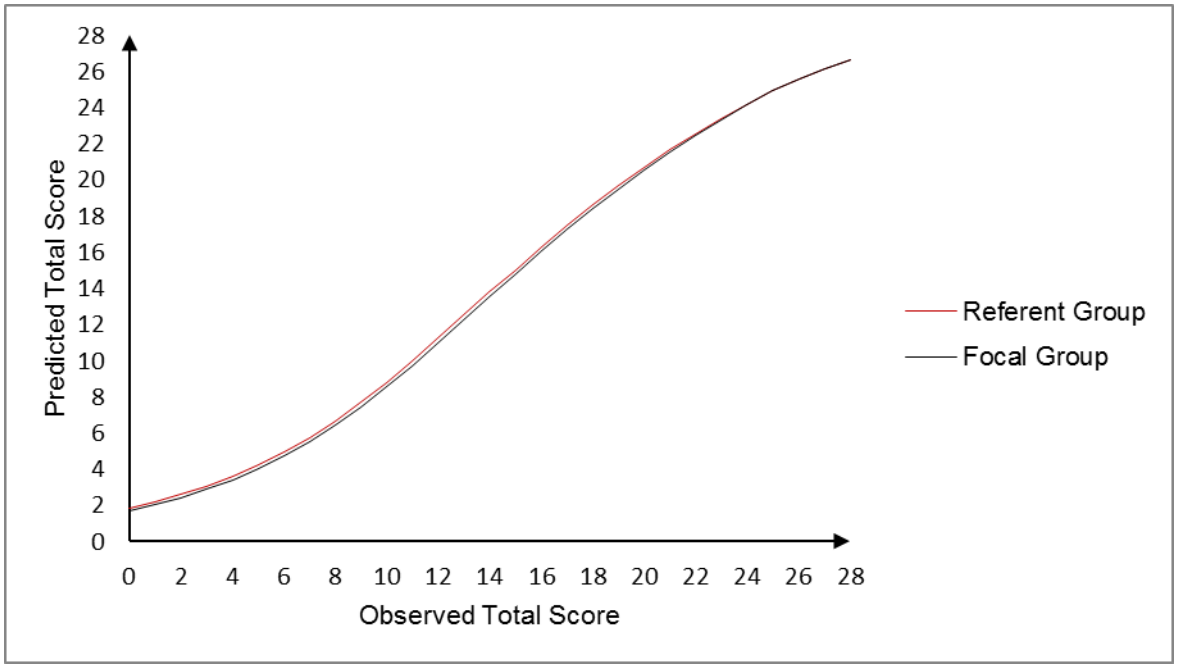

Figure 13. Test Characteristic Curves based on LR DIF parameters for White (Referent) and Non-white (Focal) groups responding to items of the SPM.

Examining the TCCs in Figure 13, the difference in predicted total score between focal and referent groups is very small, representing a mean difference of 0.16 of a raw score point in favour of the White group across all points of the curve. Even at its largest (at an observed score of 13 raw score points), this only equates to a difference of 0.26 of a raw score point favouring the White group, much smaller than the mean difference in test performance of 1.5 raw score points observed in section 3.2.3.1. It would appear, then, that the SPM does not function differently for White and Non-white groups.

However, it is possible that the conflation of all other ethnic groups to Non-white may have confounded the results of the DIF analysis above. For this reason, further ethnic DIF analyses were carried out comparing White test takers (as the referent group) with test takers from the two largest ethnic groups in the dataset: Asian test takers (N= 269) and Middle Eastern test takers (N= 172). As the LR method requires a minimum sample size within the focal (i.e. minority) group of around 200 participants (Zumbo, 1999), these were the only ethnic groups in the dataset for which further analyses could realistically be conducted.

SPM White-Asian DIF Analysis – Iteration 1

	Nagelkerke R^2 values at each step in the hierarchical regression					
	Step 1	Step 2 (uniform DIF)	Step 3 (non-uniform DIF)	DIF χ^2 (2) test	DIF R^2 (ΔR^2 between Steps 1 & 3)	DIF detected
Item 1	.361	.382	.382	5.95	.021	Negligible
Item 2	.141	.144	.144	1.08	.003	Negligible
Item 3	.149	.155	.155	1.20	.006	Negligible
Item 4	.242	.243	.244	0.42	.002	Negligible
Item 5	.234	.236	.236	1.28	.002	Negligible
Item 6	.240	.241	.244	0.94	.004	Negligible
Item 7	.317	.318	.336	9.60**	.019	Negligible
Item 8	.327	.327	.328	1.10	.001	Negligible
Item 9	.300	.304	.305	2.81	.005	Negligible
Item 10	.453	.458	.458	4.73	.005	Negligible
Item 11	.338	.351	.353	5.53	.015	Negligible
Item 12	.314	.315	.317	1.59	.003	Negligible
Item 13	.316	.317	.318	0.77	.002	Negligible
Item 14	.298	.301	.301	1.22	.003	Negligible
Item 15	.271	.271	.272	0.52	.001	Negligible
Item 16	.353	.353	.353	0.07	.000	Negligible
Item 17	.318	.318	.321	1.16	.003	Negligible
Item 18	.589	.590	.590	0.14	.001	Negligible
Item 19	.465	.465	.465	0.23	.000	Negligible
Item 20	.481	.488	.488	4.03	.007	Negligible
Item 21	.470	.471	.471	0.37	.001	Negligible
Item 22	.432	.432	.432	0.22	.000	Negligible
Item 23	.290	.290	.292	1.12	.002	Negligible
Item 24	.438	.438	.443	3.90	.005	Negligible
Item 25	.412	.417	.418	4.21	.006	Negligible
Item 26	.480	.481	.487	5.50	.007	Negligible
Item 27	.420	.420	.441	13.03**	.021	Negligible
Item 28	.384	.384	.391	4.78	.007	Negligible

Table 8. Initial iteration for LR Method identification of White-Asian ethnic DIF. Note: *$p < .05$; **$p < .01$; ***$p < .001$

As was the case for the White-Non-white analysis, some significant DIF effects were detected, though the effect sizes of these were negligible. The analysis was terminated after this iteration. As in the previous analyses, TCCs were generated for focal and referent groups to explore potential DTF effects. The resultant curves are shown in Figure 14 below:

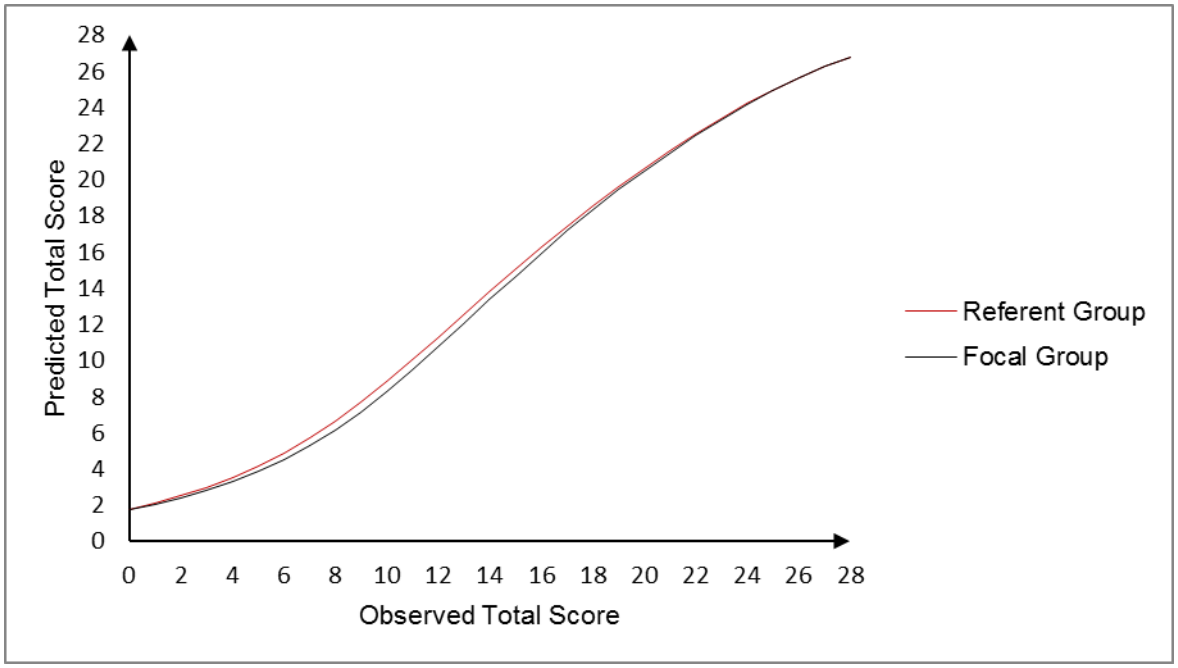

Figure 14. Test Characteristic Curves based on LR DIF parameters for White (Referent) and Asian (Focal) groups responding to items of the SPM.

Again, the TCCs shown in Figure 14 for focal and referent groups matched one another closely, representing a mean difference of 0.23 of a raw score point in favour of the White group. Though this difference is larger than that observed in the previous analysis, the largest difference between curves was still only 0.55 of a raw score point at an observed score of 11, much less than the observed mean difference of 1.07 raw score points between these groups. These results suggest that, as in the White-Non-white analysis, no substantial DTF effect could be detected between White and Asian groups.

SPM White-Middle Eastern DIF Analysis – Iteration 1

	Nagelkerke R^2 values at each step in the hierarchical regression					
	Step 1	Step 2 (uniform DIF)	Step 3 (non-uniform DIF)	DIF χ^2 (2) test	DIF R^2 (ΔR^2 between Steps 1 & 3)	DIF detected
Item 1	.277	.313	.314	8.25*	.037	Negligible
Item 2	.146	.159	.162	4.12	.016	Negligible
Item 3	.130	.133	.134	0.72	.004	Negligible
Item 4	.206	.207	.210	0.64	.004	Negligible
Item 5	.227	.227	.228	0.40	.001	Negligible
Item 6	.238	.240	.240	0.46	.002	Negligible
Item 7	.414	.414	.414	0.17	.000	Negligible
Item 8	.356	.358	.358	1.06	.002	Negligible
Item 9	.358	.363	.372	7.67*	.014	Negligible
Item 10	.481	.483	.485	2.89	.002	Negligible
Item 11	.276	.276	.277	0.15	.001	Negligible
Item 12	.223	.229	.231	2.62	.008	Negligible
Item 13	.304	.313	.319	4.95	.015	Negligible
Item 14	.259	.264	.265	1.99	.006	Negligible
Item 15	.268	.274	.275	3.88	.007	Negligible
Item 16	.321	.332	.336	8.57*	.015	Negligible
Item 17	.365	.369	.371	1.75	.006	Negligible
Item 18	.489	.519	.520	5.91	.031	Negligible
Item 19	.445	.465	.472	6.27	.027	Negligible
Item 20	.503	.504	.507	1.67	.004	Negligible
Item 21	.493	.494	.500	2.15	.007	Negligible
Item 22	.431	.431	.432	0.33	.001	Negligible
Item 23	.311	.313	.313	0.92	.002	Negligible
Item 24	.464	.465	.466	1.17	.002	Negligible
Item 25	.471	.472	.478	4.46	.007	Negligible
Item 26	.504	.505	.509	3.56	.005	Negligible
Item 27	.469	.488	.496	16.30***	.027	Negligible
Item 28	.376	.377	.391	8.49*	.025	Negligible

Table 9. Initial iteration for LR Method identification of White-Middle Eastern ethnic DIF. Note: *$p < .05$; **$p < .01$; ***$p < .001$

As for the previous analyses, some significant DIF effects were detected. However, the practical effect sizes observed for these items were negligible, so the analysis terminated after this iteration. As before, TCCs were generated based on the regression coefficients obtained in the DIF analyses. These are shown in Figure 15 below:

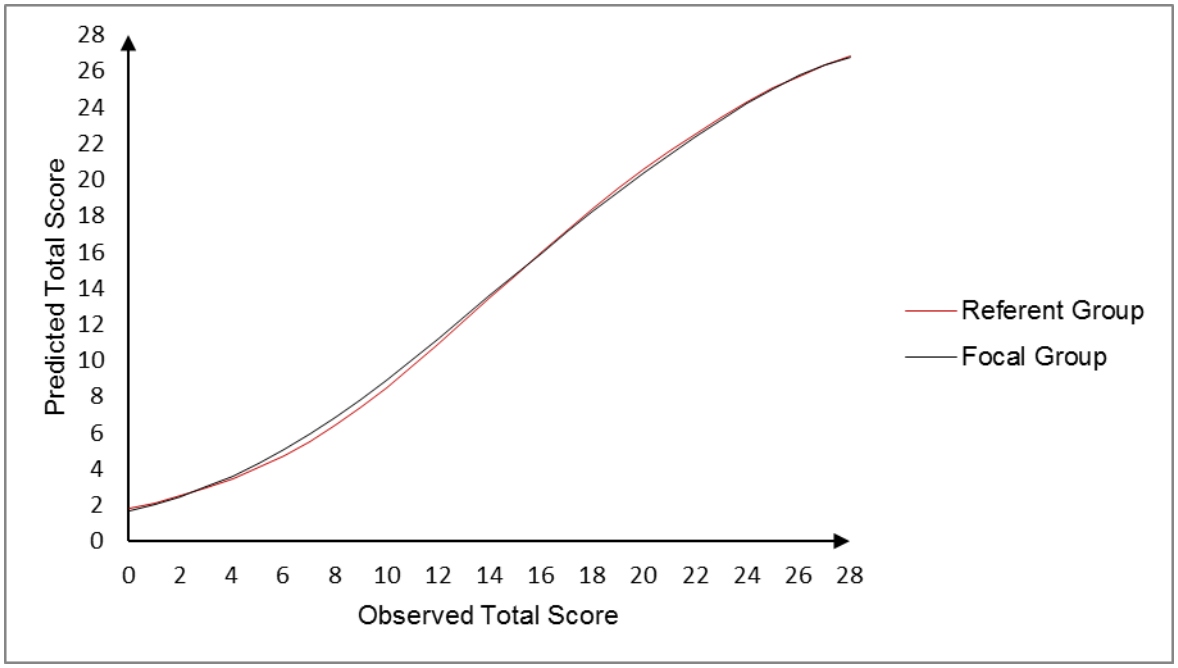

Figure 15. Test Characteristic Curves based on LR DIF parameters for White (Referent) and Middle Eastern (Focal) groups responding to items of the SPM.

As in the DTF analyses for White-Non-white and White-Asian groups, the TCCs of the focal and referent groups were very similar in form, the mean difference between the curves being 0.07 raw score points in favour of the Middle Eastern group. The largest difference between the curves was 0.48 raw score points at an observed score of 9, again in favour of the Middle Eastern group. The largest difference between the curves in favour of the White group was 0.20 raw score points at an observed score of 20. Given that the observed mean difference between these groups was 0.89 raw score points in favour of the White group, the observed DTF effect is clearly unable to account for this difference in performance.

Together, these findings would appear to support H_2, in that any of the DTF effects detected by the LR analyses are unable to explain more than a trivial proportion of the observed mean performance differences. However, it is not clear at this stage whether the LR Method is not sensitive enough to detect ethnic DTF in the dataset, or whether, in fact, no ethnic DTF exists for the SPM. To further explore this, a more sophisticated approach to the identification of DTF was employed.

3.2.3.3 DTF Analysis using MLVM

Mixture Latent Variable Modelling as an approach to the identification of ethnic DTF is relatively new. It has been adapted by bias researchers specifically for the identification of DTF by its use to identify heterogeneity within samples of the likelihood of groups responding correctly to a test's items on the basis of thresholds and factor loadings between indicators (item responses) and a single latent variable (θ) that are conditional upon membership of particular latent classes. The implication of this is that two members of opposing latent classes might have the same level of θ, but their membership within their particular class may lead them to have different probabilities of responding correctly to the items in a test, potentially leading to the test measuring their ability differently. The advantage of MLVM over traditional measures is that the nature of these latent classes is not defined *a priori*, only their number. This allows for heterogeneity within ethnic groups on whichever variable the latent classes represent, something that traditional approaches to DIF/DTF identification lack. Once these latent classes have been identified (and the assumptions of MLVM have been checked), the latent classes can be examined to determine what secondary variable most clearly defines them. Sawatzky et al. (2012) recommend that this is investigated by using the newly-created latent class membership variable as the dependent variable in a series of binary (in the case of the existence of two distinct latent classes) or multinomial (if the analysis identifies more than two distinct latent classes) logistic regressions, the independent variables within which are the variables that could potentially better explain class membership than ethnicity alone.

Statistically, MLVM analysis functions using a robust maximum likelihood estimator with robust standard errors. A number of sets of random starting values for all factor loadings and thresholds are specified, and Mplus calculates the log likelihood values of class membership on the basis of these (Stage 1). Once these have been computed, an optimisation stage (Stage 2) runs the analyses again for the starting values (the number of which is, again, specified by the user) that show the greatest differences in response

probabilities across classes. The set of starting values that displays the greatest difference in terms of likelihood between classes is then chosen to represent the parameters of the mixture model.

The interpretation of the results of MLVM is typically broken down into four stages. In the first stage, the assumption of unidimensionality is examined. As stated above, models in MLVM are specified such that all indicators load on to a single latent variable to represent ability. If, through factor analysis, it is demonstrated that more than one factor underlie the data, this potentially makes any interpretation of the results of the MLVM output null and void. In this stage, the underlying factor structure of the one-class model is assessed using factor analysis techniques to establish that the factor structure is unidimensional. Assuming that the data is unifactorial (as it should be in the case of a test of specific ability), a poor fit of the one-class model to the data indicates that the model parameters are not consistent across two or more latent classes. In the second stage, the assumption of sample homogeneity is tested. This gives an indication of how good a fit the MLVM is to the data, how well the model classifies participants into categories within the latent class variable and what the optimal number of classes is for this variable. In the third stage, the implications of sample heterogeneity are evaluated. This provides information about the implications of ignoring sample heterogeneity in terms of how the specific classes within the sample might be advantaged or disadvantaged relative to one another at each level of θ. Finally, sources of sample heterogeneity are identified using the logistic regression procedure described above.

The one-class model and three MLVMs were run in Mplus. In each model, the 28 dichotomous response variables were made to load on to a single latent factor. For each model, the number of latent classes within the class membership variable was specified. The models were specified so that the first parameter was freed, allowing factor loadings and thresholds to vary across latent classes. Mplus was configured to use 5000 sets of random starting values in stage 1 of the analysis, followed in stage 2 by the 1000 of these

that displayed the greatest log likelihood values. Though Muthén and Muthén (2010) warn that the use of this many random starting value sets will dramatically increase the computational demands of the analysis, Sawatzky et al. (2012) recommend this number for robust analysis of large datasets.

Before the output of Mplus was analysed, the latent structure of the response data was first examined in Mplus using Exploratory Factor Analysis (EFA) with PROMAX (oblique) rotation and unweighted least squares (ULSMV) estimation, as this method has been shown to perform better than alternative extraction methods when used on binary data (Parry & McArdle, 1991). All indicators were flagged as categorical to enable the generation of tetrachoric correlations. Conducting the EFA in this manner ensured that the binary response data were analysed properly, something that is known to be problematic for EFA procedures conducted in the current iteration of SPSS (Tran & Formann, 2009), at least without the use of complex specialist macros.

The eigenvalue of the first extracted factor was 12.13. However, there were an additional 4 factors with eigenvalues over 1, indicating a 5-factor solution by the Kaiser criterion. The scree plot of eigenvalues obtained is shown in Figure 16 below:

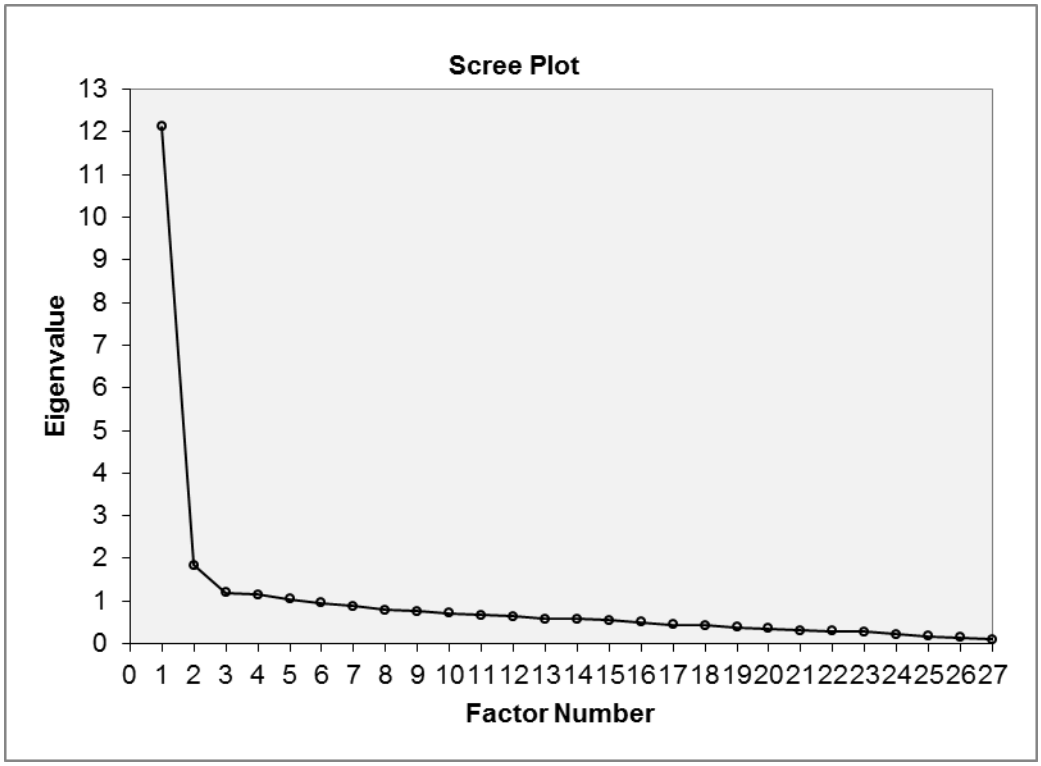

Figure 16. Scree Plot of eigenvalues for factors underlying the SPM.

Examining Figure 16, the scree plot shows a very sharp point of inflexion at the 3rd factor. This indicates that two factors underlie the data. However, while more reliable than the Kaiser Criterion, the scree test is, by its nature, a subjective measure, so it is common practice to confirm the factor solution to which it points using a more reliable and objective method. A relatively reliable – and extremely widely-used – statistical method for determining the correct number of factors underlying the data would be to conduct parallel analysis on it according to the procedure recommended by O'Connor (2000). However, Tran and Formann (2009) have noted that traditional parallel analysis methods tend to provide unreliable results with binary data. Thankfully, there are a number of modern alternatives for dimension identification, based on tetrachoric correlations, which have been demonstrated to perform well in simulations (Timmerman & Lorenzo-Seva, 2011). The freeware program FACTOR (Lorenzo-Seva & Ferrando, 2006) is able to perform many of these forms of parallel analysis on binary and ordered polytomous data. FACTOR v10.3.01

64bits for Windows was used to conduct parallel analysis based on minimum rank factor analysis (PA-MRFA; Timmerman & Lorenzo-Seva, 2011). This analysis compares tetrachoric correlation matrices from the data with 500 random correlation matrices based on permutations of the raw data (Buja & Eyuboglu, 1992). The analysis indicated two factors underlying the data (with eigenvalues of 12.13 and 1.84), jointly explaining 37.5% of the variance.

This was somewhat problematic, as MLVM depends on the assumption of a single factor underlying the data. At this point, it was unclear whether or not the second extracted factor represented a relevant additional substantive latent factor (Sawatzky et al., 2012), and whether – by extension – this represented a true multifactorial underlying structure to the SPM. Without being sure of this, any potential ethnic DTF detected might, instead, be due to differing levels between groups of traits associated with these secondary factors. To investigate this further, the obtained factor solution was carefully interpreted to determine whether the second factor truly represented a substantive latent factor (see section 3.4.2).

Nevertheless, the analysis continued by trying to establish whether sample heterogeneity could be identified in the responses. This was done by comparing the fit statistics of competing models. Mplus calculates three measures of model fit by which the correct number of latent classes can be adjudged: the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the sample-adjusted BIC. For all of these statistics, a lower value indicates better model fit, the lowest value of all models being the one that the criterion judges to be the most likely number of latent classes. These statistics are comparative in nature, meaning that the absolute value of them for a model tells us little in isolation. However, when comparing models, a difference of only 10 between BIC values represents odds of 150:1 that the model with the lower value represents poorer model fit than the model with the higher one (Raftery, 1995).

Monte Carlo simulations (e.g. Nylund, Asparouhov, & Muthén, 2007) have demonstrated that the BIC provides a more reliable measure of the correct number of latent classes within a dataset than the AIC does. Furthermore, Yang (2006) has demonstrated that the procedure used to calculate the sample-adjusted BIC demonstrates substantial improvement to estimations of number of latent classes than the unadjusted BIC. For these reasons, the correct number of latent classes was judged using the sample-adjusted BIC.

Table 10 shows the fit statistics generated by Mplus for the models tested:

	AIC	**BIC**	**Sample-adjusted BIC**	**Entropy**
One class	21937.13	**22214.91**	22037.04	-
Two-class	21833.04	22393.56	**22034.66**	.62
Three-class		Unidentified		
Four-class	**21777.37**	22903.37	22182.38	.73

Table 10. Fit statistics generated by Mplus for the 1-, 2-, 3- and 4-class models.

Based on the sample-adjusted BIC, data fit the 2-class model better than any other model. The entropy statistic of .62 observed for the 2-class model indicates that participants can be categorised well as belonging to one of two latent classes according to their pattern of response behaviour. However, this finding was not supported by the differences observed in the AIC and BIC statistic. Though the sample-adjusted BIC has been demonstrated to be a better criterion with which to judge the correct number of classes for MLVM (Yang, 2006), it would be foolish not to examine some of the alternative models in further analyses, given the lack of clarity in the results above. The 3-class model was unidentified, suggesting that heterogeneity in the data could not be explained by the existence of three distinct patterns of response behaviour. The 4-class model, however, demonstrated the lowest value on the AIC criterion. Furthermore, the 4-class model's entropy value of .73 suggests that this model classified participants into latent classes with somewhat less overlap than the 2-class

model did. This suggests the possibility that, in fact, four distinct classes of response pattern underlie the data.

To clarify these findings, the parameters generated for each item in the MLVM were inspected for both the 2-class and 4-class models, and were compared to those generated for the One-class model. These parameters are shown in Tables 11 and 12 below.

Item	One-class Model			Class 1			Class 2		
	λ	τ	b	λ	τ	b	λ	τ	b
1	1.61	-3.46	-2.15	1.33	-3.58	-2.69	2.34	-3.10	-1.33
2	0.93	-2.71	-2.90	0.81	-2.65	-3.29	1.41	-3.00	-2.13
3	0.95	-3.64	-3.81	0.34	-3.88	-11.38	1.67	-3.14	-1.88
4	1.33	-3.67	-2.75	0.94	-3.63	-3.88	2.29	-3.79	-1.66
5	1.08	-1.74	-1.60	1.12	-1.85	-1.65	0.90	-1.36	-1.51
6	1.19	-3.15	-2.66	0.94	-3.00	-3.20	2.20	-4.03	-1.83
7	1.57	-2.04	-1.30	1.41	-2.25	-1.60	2.38	-1.44	-0.60
8	1.49	-1.59	-1.07	1.43	-1.74	-1.21	1.52	-1.04	-0.69
9	1.28	-1.49	-1.17	1.19	-1.46	-1.24	1.65	-1.64	-0.99
10	1.75	-0.10	-0.06	1.76	-0.23	-0.13	1.81	0.42	0.23
11	1.40	-3.09	-2.21	1.20	-3.43	-2.86	2.00	-2.34	-1.17
12	1.23	-2.44	-1.99	1.08	-2.45	-2.27	1.68	-2.40	-1.43
13	1.53	-2.85	-1.87	1.65	-3.49	-2.12	1.55	-1.67	-1.08
14	1.19	-2.20	-1.85	1.05	-2.30	-2.18	1.61	-1.86	-1.16
15	0.85	-0.31	-0.37	0.78	-0.40	-0.51	1.04	0.03	0.03
16	1.15	0.33	0.29	1.15	0.13	0.11	1.06	1.09	1.03
17	1.58	-3.17	-2.00	1.36	-3.10	-2.29	2.46	-3.55	-1.44
18	2.98	-5.56	-1.86	3.00	-5.61	-1.87	3.24	-5.81	-1.80
19	2.38	-4.39	-1.84	2.31	-4.37	-1.89	2.69	-4.58	-1.71
20	2.36	-3.05	-1.29	2.02	-2.71	-1.34	6.01	-7.03	-1.17
21	2.49	-3.77	-1.51	2.42	-3.60	-1.49	3.56	-5.45	-1.53
22	2.01	-2.26	-1.13	1.87	-1.95	-1.04	7.02	-9.07	-1.29
23	1.20	-1.62	-1.35	1.42	-1.80	-1.27	0.58	-1.16	-2.01
24	1.83	-1.29	-0.71	2.30	-1.09	-0.48	1.89	-2.92	-1.54
25	1.81	-1.17	-0.65	2.10	-0.87	-0.41	3.81	-4.98	-1.31
26	1.76	0.43	0.25	2.34	0.28	0.12	0.46	0.84	1.81
27	1.93	2.11	1.09	2.67	2.12	0.79	-0.82	3.32	-4.07
28	1.25	1.36	1.08	1.48	1.26	0.85	0.21	1.72	8.28

Table 11. MLVM parameters generated for the SPM (2-class solution).

Examining Table 11, the factor loadings (λ) and thresholds (τ) for each class in the 2-class model all look relatively sensible. Conversely, in Table 12, many of the factor loadings and thresholds in the 4-class model appear very large, particularly in the case of Class 4. This suggests that this class may be a statistical artefact, and may not be representative of a true

subpopulation within the dataset. On the basis of the item parameters for each of the two models under consideration, then, the 2-class model would appear to be the more defensible solution.

Item	One-class Model			Class 1			Class 2			Class 3			Class 4		
	λ	τ	*b*	λ	τ	*b*	λ	τ	*b*	λ	τ	*b*	λ	τ	*b*
1	1.61	-3.46	-2.15	2.33	-2.85	-1.22	1.57	-4.07	-2.59	0.76	-2.18	-2.85	3.39	-4.35	-1.29
2	0.93	-2.71	-2.90	1.01	-2.20	-2.17	1.74	-4.22	-2.43	0.41	-1.34	-3.28	-0.86	-0.91	1.07
3	0.95	-3.64	-3.81	1.23	-2.54	-2.07	1.09	-4.68	-4.28	-0.75	-3.25	4.37	1.07	-2.59	-2.41
4	1.33	-3.67	-2.75	1.71	-2.92	-1.71	1.65	-4.44	-2.69	-3.59	-8.08	2.25	0.22	-1.81	-8.34
5	1.08	-1.74	-1.60	1.65	-1.68	-1.02	1.10	-1.85	-1.69	1.38	-1.17	-0.85	-0.33	-27.36	84.24
6	1.19	-3.15	-2.66	6.02	-9.08	-1.51	1.09	-3.12	-2.87	-0.17	-2.17	13.10	0.64	-2.93	-4.58
7	1.57	-2.04	-1.30	1.84	-1.19	-0.65	1.59	-2.49	-1.56	9.07	-6.95	-0.77	0.75	-0.52	-0.69
8	1.49	-1.59	-1.07	2.18	-1.71	-0.79	1.49	-1.87	-1.26	0.59	-0.74	-1.26	1.02	0.33	0.32
9	1.28	-1.49	-1.17	1.52	-1.65	-1.09	1.13	-1.49	-1.32	1.02	-1.87	-1.83	93.86	35.17	0.38
10	1.75	-0.10	-0.06	3.20	-0.11	-0.04	1.44	-0.18	-0.13	1.81	0.04	0.02	17.53	11.57	0.66
11	1.40	-3.09	-2.21	2.38	-2.57	-1.08	1.07	-3.28	-3.07	0.08	-2.89	-35.57	9004.30	-6171.39	-0.69
12	1.23	-2.44	-1.99	0.84	-1.37	-1.63	1.52	-3.76	-2.48	1.29	-0.83	-0.64	8.00	-1.82	-0.23
13	1.53	-2.85	-1.87	1.78	-1.69	-0.95	1.47	-3.25	-2.21	0.96	-3.30	-3.44	8987.35	-6159.65	-0.69
14	1.19	-2.20	-1.85	2.33	-2.64	-1.14	1.00	-2.37	-2.37	-0.40	-1.56	3.92	538.08	-40.99	-0.08
15	0.85	-0.31	-0.37	0.62	0.13	0.20	0.99	-0.57	-0.58	0.05	0.43	9.18	1.99	0.01	0.00
16	1.15	0.33	0.29	2.69	0.62	0.23	0.96	0.16	0.17	0.25	1.51	6.11	2.60	-0.11	-0.04
17	1.58	-3.17	-2.00	2.79	-4.45	-1.60	1.35	-2.95	-2.19	-1.04	-3.98	3.84	6.87	-4.32	-0.63
18	2.98	-5.56	-1.86	2.42	-4.94	-2.04	3.60	-6.49	-1.81	0.85	-4.03	-4.75	4.54	-3.84	-0.85
19	2.38	-4.39	-1.84	1.18	-2.78	-2.37	3.06	-5.43	-1.77	4.99	-11.72	-2.35	8987.50	-6159.76	-0.69
20	2.36	-3.05	-1.29	9.99	-11.37	-1.14	2.73	-3.29	-1.21	-0.64	-2.04	3.17	265.10	-212.70	-0.80
21	2.49	-3.77	-1.51	3.70	-5.13	-1.39	2.54	-3.90	-1.53	1.26	-2.38	-1.88	9.91	-10.65	-1.07
22	2.01	-2.26	-1.13	13.69	-18.02	-1.32	2.04	-2.31	-1.13	0.81	-0.81	-1.01	4.66	-1.63	-0.35
23	1.20	-1.62	-1.35	1.33	-2.27	-1.71	1.29	-1.71	-1.32	0.45	-0.90	-1.99	0.86	-0.20	-0.23
24	1.83	-1.29	-0.71	2.32	-3.52	-1.52	2.07	-1.35	-0.65	3.73	0.01	0.00	0.46	0.43	0.94
25	1.81	-1.17	-0.65	5.67	-8.37	-1.48	2.00	-1.02	-0.51	0.91	-0.34	-0.38	16.42	4.21	0.26
26	1.76	0.43	0.25	1.72	0.06	0.04	1.65	0.47	0.28	36.79	16.98	0.46	2.88	-0.23	-0.08
27	1.93	2.11	1.09	2.12	0.89	0.42	5.09	5.68	1.11	2.81	1.33	0.47	0.44	1.01	2.32
28	1.25	1.36	1.08	2.06	0.67	0.32	1.51	1.91	1.26	0.75	0.29	0.39	0.82	1.25	1.53

Table 12. MLVM parameters generated for the SPM (4-class solution).

Once a defensible solution has been identified, the next step in Sawatzky et al.'s (2012) process is to evaluate the implications of heterogeneity. One approach to examining the implications of sample heterogeneity visually in an easily interpretable way is to plot TCCs for each class and compare them to that of the one-class model. Following the logic of the procedure described by Oliveri et al. (2012), for each level of θ, the logit for each item may be calculated by subtracting the item's difficulty parameter (b) from θ, and multiplying the result by its discrimination parameter (λ). From these values, predicted item responses may be generated for each item at a range of levels of θ to generate an ICC for that item. These ICCs can then be summated to produce a TCC.

TCCs were generated for each class and for the total sample, based on the parameters shown in Table 11. The logit for each item was calculated based on a range of θ derived from the maximum and minimum factor scores generated for the one-class model and as part of the MLVM analyses. This gave a range of θ from -3 to 3, which were then used to calculate each item's logit in increments of 0.1 of an S.D. unit (to ensure that the curves generated were smooth). Figure 17 below shows the resultant TCCs for each latent class and for the one-class model:

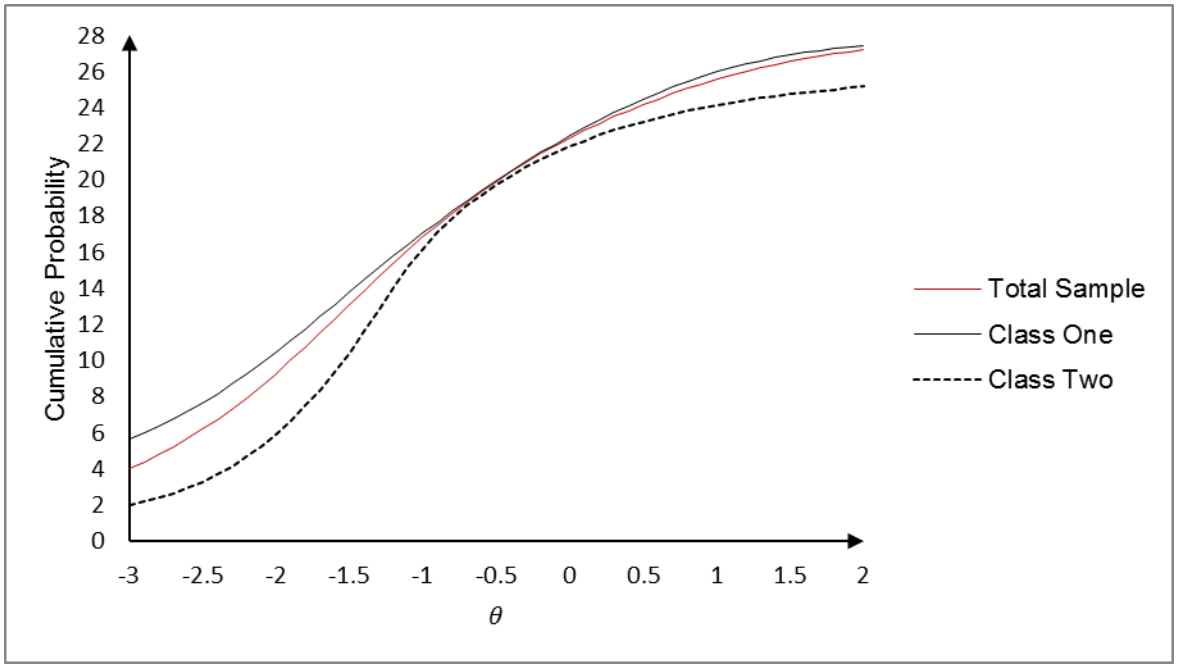

Figure 17. TCCs for each latent class in the 2-class model and for the total sample (One-class model).

Examining Figure 17, it is clear that there are non-trivial implications for ignoring sample heterogeneity. Though little difference between the curves can be seen at moderate levels of θ (from around -1 to 0), at other levels – particularly at very high and very low levels of θ – the test appears to consistently favour members of Class One over members of Class Two. This observable difference in TCCs represents a real, non-ignorable DTF effect.

Having established that non-trivial DTF effects exist, the final step is to investigate the potential sources of sample heterogeneity, principally to examine to what extent this effect is representative of ethnic DTF. To this end, the classes generated in the MLVM analyses were scrutinised more closely. Summary statistics for classes within the 2-class model are shown in Table 13 below:

	N	Mean SPM Raw Score (S.D.)	*z*-score
Class 1	856	21.67 (4.76)	0.21
Class 2	198	19.52 (5.04)	-0.23

Table 13. Number, mean raw score and z-score for each class within the 2-class model.

Examining Table 13, there is evidence that differences in response pattern lead to real, measureable test performance between classes that appear not to depend on cognitive ability. In the 2-class model, there is a difference of approximately 0.4 S.D. units between classes 1 and 2.

At first inspection, then, these results appear to parallel aspects of the ethnic group mean score differences observed in previous analyses, as well as those observed in the literature between ethnic groups in terms of their mean performance. The next logical step in the analyses would, therefore, be to examine the degree to which these latent classes matched manifest ethnic groups. A breakdown of the frequency of ethnic groups within each class is shown in Tables 14, 15 and 16 below:

	White N (% of group)	**Non-white N (% of group)**
Class 1	362 (82.1%)	493 (80.6%)
Class 2	79 (17.9%)	119 (19.4%)

Table 14. Frequency of White and Non-white participants within each latent class.

	White N (% of group)	**Asian N (% of group)**	**Black N (% of group)**	**Middle Eastern N (% of group)**	**Latin-American N (% of group)**
Class 1	362 (82.1%)	213 (79.2%)	16 (66.7%)	139 (80.8%)	1 (100%)
Class 2	79 (17.9%)	56 (20.8%)	8 (33.3%)	33 (19.2%)	0 (0%)

Table 15. Distribution of broad ethnic groups within each latent class.

	White – British N (% of group)	**Other White N (% of group)**	**Indian N (% of group)**	**Pakistani N (% of group)**	**Other Asian N (% of group)**	**Middle Eastern N (% of group)**
Class 1	206 (82.1%)	140 (83.3%)	98 (74.8%)	42 (89.4%)	58 (82.9%)	139 (80.8%)
Class 2	45 (17.9%)	28 (16.7%)	33 (25.2%)	5 (10.6%)	12 (17.1%)	33 (19.2%)

Table 16. Distribution of reduced ethnic categories within each latent class.

Examining these tables, there seem to be similar class distributions across ethnic groups. Most ethnic groups appear to have equivalent proportions of their total within each latent

class. This would appear to suggest that ethnicity is largely orthogonal to latent class membership.

To confirm these observations, Sawatzky et al. (2012) recommend using the newly-created class variable as the dependent variable in a series of logistic regression analyses, the predictors being ethnic group variables within the dataset. Significant prediction of group membership by ethnicity variables in the logistic regression model is indicative of a degree of DTF between ethnic groups, even if the pseudovariance explained by these predictors is fairly minimal. If identified, this ethnic DTF can then be explored by examining the secondary variables that are associated more strongly with class membership.

Latent class membership was entered into a binary logistic regression model as the dependent variable. In the first model (shown in Table 17), ethnicity classified as White-Non-white was entered as the predictor. In the second model (shown in Table 18), broad ethnic group was entered as a categorical predictor (the Latin-American category having been removed). In the third model (shown in Table 19), reduced ethnic classification was entered as a categorical predictor. In the latter two models, the first category in the ethnicity variable (White and White – British respectively) was used as the reference category. The output for these analyses is shown below.

		95% CI for Odds Ratio		
	B (SE)	***Lower***	***Odds Ratio***	***Upper***
Constant	-1.52 (0.12)			
White-Non-white	0.10 (0.16)	0.81	1.11	1.52

Table 17. Logistic regression for White-Non-white classification predicting latent class in the 2-class model. Note: R^2 = .000 (Cox & Snell), .001 (Nagelkerke). Model $\chi^2(1) = 0.40$, $p = .53$.

	B (SE)	95% CI for Odds Ratio		
		Lower	*Odds Ratio*	*Upper*
Constant	-1.44			
White	(Reference)			
Asian	-0.08 (0.23)	0.59	0.92	1.44
Black	0.10 (0.25)	0.69	1.11	1.79
Middle Eastern	0.75 (0.47)	0.83	2.11	5.34

Table 18. Logistic regression for broad ethnic classification (Latin-American excluded) predicting latent class in the 2-class model. Note: R^2 = .00 (Cox & Snell), .01 (Nagelkerke). Model χ^2 (3) = 3.57, p = .31.

	B (SE)	95% CI for Odds Ratio		
		Lower	*Odds Ratio*	*Upper*
Constant	-2.13 (0.47)			
White – British	(Reference)			
Other White	0.61 (0.50)	0.69	1.84	4.90
Indian	0.52 (0.52)	0.61	1.68	4.62
Pakistani	0.55 (0.57)	0.57	1.74	5.31
Other Asian	0.69 (0.51)	0.73	1.99	5.43
Middle Eastern	1.04 (0.51)	0.04	2.83	7.75

Table 19. Logistic regression for reduced ethnic classification predicting latent class in the 2-class model. Note: R^2 = .01 (Cox & Snell), .01 (Nagelkerke). Model χ^2 (5) = 6.36, p = .27

For the 2-class model, none of the three representations of ethnicity was found to be a predictor of group classification. The third model explained slightly more of the pseudovariance than the two preceding models, though this did not represent adequate explanation of class membership. The results of these analyses suggest that ethnicity cannot satisfactorily explain class membership for the 2-class model.

Considering these findings alongside those of the LR DTF analysis, the totality of evidence would suggest that, in the case of the current dataset, the Raven's SPM does not appear to produce ethnic DTF. These findings do not support H_3.

The final point to address in these analyses, then, is to attempt to interpret the obtained latent classes. The remaining demographic variables were used as predictors of latent class membership in a series of binary logistic regression models, but none of them were able to explain more than trivial pseudovariance in class membership (Gender: B = -0.01, p = .98; Age: B = -0.02, p = .65; Educational Level: B = -0.06, p = .30).

Since none of the demographic variables in the dataset can account for latent class membership, an alternative approach is needed that considers, at the item level, what might characterise membership of each of the classes. To do this, the obtained thresholds for each latent class may be used to derive odds ratios for each item. These odds ratios represent the odds of a correct item response in Class One relative to the odds of a correct response in Class Two. The odds ratios generated for each item are shown in Table 20.

Item	Odds Ratios Class 1 vs. Class 2
1	1.62
2	0.70
3	2.10
4	0.85
5	1.63
6	0.36
7	2.25
8	2.01
9	0.84
10	1.92
11	2.97
12	1.05
13	6.17
14	1.55
15	1.54
16	2.61
17	0.64
18	0.82
19	0.81
20	0.01
21	0.16
22	0.001
23	1.90
24	0.16
25	0.02
26	1.75
27	3.32
28	1.58

Table 20. Odds ratios comparing odds for each latent class of a correct response to each item in the SPM. Note: Odds ratios greater than 1 indicate higher odds of a correct response for the first class in each comparison.

These odds ratios can then be used to identify the items that separate the groups in terms of their response behaviour. The magnitude of each odds ratio is first classified according to its effect size. Derived from Cohen's (1988) effect sizes for *r*, critical values of 1.49, 3.45, and 9.0 (Haddock, Rindskopf & Shadish, 1998) are used to classify odds ratios as representing small, medium, and large effects respectively for odds ratios greater than 1 (i.e. favouring Class One). For odds ratios less than 1 (i.e. favouring Class Two), the reciprocal vales of 0.67, 0.29 and 0.11 are used to denote small, medium, and large effect sizes.

Examining Table 20, the majority of the odds ratios obtained represent either small or trivial effect sizes. The majority of these favour Class One, as is to be expected, based on this class' better performance relative to Class Two. However, there are a handful of these effects that are classified as either medium or large, namely Item 15's effect (favouring Class One), and those of Items 20, 21, 22, 24 and 25 (all of which favour Class Two). These appear to be the items that would most likely be able to characterise the latent classes.

The challenge, then, is to determine what intrinsic property of these items leads members of the latent classes to respond differently to them. For MLVMs conducted on data in health and related settings (e.g. Sawatzky et al., 2012), this is relatively simple, as the indicators themselves are often descriptive of particular medical conditions, attitudes, and so on. This allows latent classes to be characterised by examination of the content of the items, in much the same way as one might do for exploratory factor analysis. While this approach can be relatively straightforward for items contained within measures of verbal and numerical reasoning, it is much more challenging to do for measures of abstract reasoning, as the salient properties of these items are much more difficult to extract from their examination.

One observation from Table 20 is that there is a general pattern within the obtained odds ratios. It seems – though this is perhaps not as clear cut as it could be – that members of Class One are more likely than Class Two to respond correctly to the earlier items, and members of Class Two are more likely to respond correctly to the later items. This implies that the later items share some common property that is somehow absent from the earlier items (see section 3.4.2 for discussion of what these properties might represent).

3.2.3.4 MLVM on Unifactorial Item Subsets within the SPM

The EFA in the previous section suggested that underlying structure of SPM might be multifactorial, presenting a problem for MLVM in that it assumes all items within the model load on to a single latent factor. Hunter and Schmidt (2000) cite violations of the assumption

of unidimensionality in this way to be one of the three most common methodological problems with DIF/DTF research.

However, there is, potentially, a way of overcoming this limitation. A potential workaround if multidimensionality is discovered for a test would be to consider the items that load most strongly on to each factor as comprising a series of subtests. This approach is not one that has previously been given much attention in the literature before, but there is no particular reason why this approach might be incorrect: Computer Adaptive Tests make inferences about candidates' likely level of ability on the basis of dynamic item sets. Dependent on their performance, two test takers may complete the same test made up of entirely different items in entirely different quantities, but these measures are still seen as robust assessments of ability. By examining the items within a multifactorial test as subtests in their own right, follow-up MLVM can be conducted, potentially to isolate any DTF effects.

In attempting to overcome the multifactorial nature of the SPM, the first step was to establish the items that would make up the two unifactorial subtests for further analyses. The PROMAX-rotated factor matrix for the two-factor solution is shown in Table 21.

	Factor	
	1	**2**
Item 1	**.69**	
Item 2	**.47**	
Item 3	**.73**	-.27
Item 4	**.76**	
Item 5	.28	**.31**
Item 6	**.47**	
Item 7	**.52**	.23
Item 8	**.49**	.25
Item 9	**.42**	.25
Item 10	.35	.44
Item 11	**.71**	
Item 12	**.60**	
Item 13	**.65**	
Item 14	**.59**	
Item 15	**.42**	
Item 16	.32	.30
Item 17	**.57**	
Item 18	.54	.41
Item 19	**.60**	.29
Item 20	.59	.31
Item 21	.52	.39
Item 22	.44	.40
Item 23	.30	**.35**
Item 24		**.68**
Item 25	.22	**.59**
Item 26		**.68**
Item 27	-.22	**.96**
Item 28		**.64**

Table 21. Rotated factor matrix for the 2-factor solution underlying the SPM. Note: Factor loadings below .20 have been suppressed.

Examining this rotated factor matrix, two item sets can be generated for subtests that make up the SPM. Items were selected for each subtest based on two criteria. Firstly, an item's loading on a factor had to be at least .30 or greater. Secondly, the item had not to be cross-loaded on any other factor with a loading of .30 or stronger. This critical value of .30 was chosen as this is the lowest value for practical significance cited by Hair et al. (1998). For each of the two factors, the relevant factor loadings for items that meet both of these criteria are picked out in bold in Table 21 above.

Factor 1

It made most sense, in the first instance, to consider the items that load most strongly on to the first factor in the EFA. This, theoretically, is the factor that is most likely to represent ability. Fifteen of the 28 items in the SPM met the criteria described above to be selected for the Factor 1 subtest. These items were Items 1, 2, 3, 4, 6, 7, 8, 9, 11, 12, 13, 14, 15, 17, and 19.

A MLVM was then run on this reduced item set. As before, fit statistics for the one-, two-, three-, and four-class models were generated in Mplus. These are shown in Table 22.

	AIC	BIC	Sample-adjusted BIC	Entropy
One class	10252.39	**10401.17**	**10305.88**	-
Two-class	10245.78	10548.30	10354.56	.62
Three-class	**10244.71**	10854.71	10464.05	.74
Four-class	**10244.71**	10854.71	10464.05	.74

Table 22. Fit statistics generated using a reduced set of SPM items based on loadings on the first extracted factor.

Examining these fit statistics, the values of BIC and sample-adjusted BIC are lowest for the one-class model. There is no improvement in fit statistics for more complex models, apart from a very marginal improvement in AIC for the three- and four-class models. Interestingly, both of these models showed exactly the same figures for all fit statistics and entropy, implying that one (or possibly both) of these solutions is spurious. This is largely academic, however, as the much poorer BIC and sample-adjusted BIC figures for these models compared to those of the one-class model suggest that the one-class model is much more defensible. These findings would suggest that, within this subtest of the SPM, no DTF can be identified.

Factor 2

As no DTF was identified for the Factor 1 subtest, the possibility existed that the observed DTF in the main analysis might have been due to DTF within the other subtest within the SPM.

The second subtest was made up of 7 items. The items selected for this subtest were Items 5, 23, 24, 25, 26, 27, and 28. A MLVM model was run on this item set. The fit statistics generated in Mplus are shown in Table 23 below.

	AIC	BIC	Sample-adjusted BIC	Entropy
One class	7505.31	**7574.74**	7530.28	-
Two-class	**7464.55**	7608.38	**7516.27**	.45
Three-class	7495.74	7713.96	7574.21	.38
Four-class	7522.79	7815.39	7628.00	.78

Table 23. Fit statistics generated using a reduced set of SPM items based on loadings on the second extracted factor.

Examining these fit statistics, the BIC indicates that the one-class model is the best fit to the data. The AIC and sample-adjusted BIC indicate better fit of the two-class model, though the entropy value for this model was poor. As for the Factor 1 subtest, this would suggest that no DTF was present in the Factor 2 subtest.

The implication of these findings is that there is no DTF on item subsets based on factors within the SPM. This can be viewed as additional evidence that the DTF effect identified in the previous analysis is, in actual fact, based on spurious classes due to model misspecification, rather than real response pattern differences.

3.3 Attempted Replication using Raven's APM Data Archive

An identical set of analyses on the second dataset were conducted with the aim of replicating the findings of the previous analyses. The second study used a different

measure to the first, namely the Raven's Advanced Progressive Matrices (APM). The second archive was very similar in nature to the first in that data were collected for the same key variables based on a global sample of test takers. However, the ethnic and national make-up of the second archive was substantially different to that of the first. It was reasoned that, if similar patterns of results could be observed for test scores on very similar measure of abstract reasoning ability across different ethnic and national groups, it could provide further support for the conclusion that while ethnic group test score differences exist for these measures, they cannot adequately be explained by DTF.

3.3.1 Method

3.3.1.1 Measures

Raven's Advanced Progressive Matrices

The measure used for this study, the APM, is part of the Raven's series of tests. It is very similar in form to the SPM, with some notable exceptions. It consists of 23 questions (again with three practice questions beforehand). The time limit for the APM is 42 minutes, 5 minutes less than that of the SPM. The main difference between it and the SPM is that the APM is much more difficult. It is primarily used to assess those expected to be at the upper end of the cognitive ability distribution, being recommended for the assessment of candidates for senior management positions and high-level consultancy roles, especially those of a highly technical nature (Talent Lens, 2009).

Demographic Variables

Having been collected by Talent Lens' online assessment platform, the variables recorded in the dataset were all exactly the same as those for the SPM dataset.

3.3.1.2 Sample

The second archive contained data from 1601 participants. Of the 1578 (98.0%) participants who reported their gender, 1172 were male (73.2%) and 406 were female (24.5%). As in the previous study, the dataset was made up of a global sample, comprising participants from 76 countries, the most frequently reported of which were the UK (579; 36.2%), Norway (126; 7.9%), Sweden (110; 6.9%), the USA (79; 4.9%), China (56; 3.5%), India (54, 3.4%), Germany (53; 3.3%), and Portugal (42; 2.6%). The ethnic breakdown of the dataset consisted of 559 participants who reported their ethnicity as 'White – British' (34.9%), 25 as 'White – Irish' (1.6%), 613 as 'Any other White background' (38.3%), 15 as 'Black or Black British – African' (0.9%), 2 as 'Black or Black British – Caribbean' (0.1%), 80 as 'Asian or Asian British – Indian' (5.0%), 17 as 'Asian or Asian British – Pakistani' (1.1%), 88 as 'Asian or Asian British – Chinese' (5.5%), one as 'Asian or Asian British – Bangladeshi' (0.1%), 52 as 'Asian or Asian British – Any other Asian background' (3.2%), 18 as 'Middle Eastern' (1.1%), 4 as 'Latin-American' (0.2%), 15 as 'Mixed – White and Asian' (0.9%), 4 as 'Mixed – White and Black Caribbean' (0.2%), 4 as 'Mixed – White and Black African' (0.2%), 15 as 'Mixed – White and Asian' (0.9%), 5 as 'Any other mixed background' (0.3%), 4 as 'Other ethnic group (not specified)' (0.2%), and 93 (5.8%) who did not report their ethnicity. The most commonly reported occupational backgrounds of participants in the dataset were accountancy (152; 9.5%), students (167; 10.4%), consultancy (132; 8.2%), IT (105; 6.6%), and finance (91; 5.7%), again representing a wide range of occupations.

3.3.1.2.1 Reclassifying Ethnicity

Ethnicity was classified according to the same strategies used in the previous study. In the reduced ethnicity classification variable, 5 ethnic categories fulfilled the inclusion criteria: White – British, Any other White background, Indian, Chinese, and Any other Asian background.

3.3.2 Results

3.3.2.1 Ethnic Group Test Score Differences

Mean group test scores, equivalent percentile scores (referenced against Pearson's International Graduates Job Applicants norm group for the APM) and Cohen's *d* were calculated for each of the three conceptualisations of ethnicity used in the study. These are shown in the following tables:

Ethnic Group	N	Mean APM Raw Score (S.D.)	Percentile	Cohen's *d* relative to White group
White	1197	13.23 (4.36)	74th	-
Non-white	310	13.17 (4.44)	74th	0.01

Table 24. Mean raw score, percentiles and Cohen's *d* by ethnicity classified as White/Non-white.

Ethnic Group	N	Mean APM Raw Score (S.D.)	Percentile	Cohen's *d* relative to White group
White	1197	13.23 (4.36)	74th	-
Asian	238	13.47 (4.52)	74th	-0.05
Black	18	11.33 (4.31)	56th	0.44
Middle Eastern	18	10.61 (4.87)	56th	0.57
Latin-American	4	13.21 (4.40)	74th	0.01

Table 25. Mean raw score, percentiles and Cohen's *d* by broad ethnic group classification.

Ethnic Group	N	Mean APM Raw Score (S.D.)	Percentile	Cohen's *d* relative to White – British group
White – British	559	12.62 (4.17)	74th	-
Other White	613	13.87 (4.45)	80th	-0.29
Indian	80	11.83 (4.45)	67th	0.18
Chinese	88	15.42 (4.15)	86th	-0.67
Other Asian	52	13.40 (4.38)	74th	-0.18

Table 26. Mean raw score, percentiles and Cohen's *d* by reduced ethnic group classification.

The data in Tables 24, 25 and 26 indicate that there are group differences in mean test score across ethnic groups, though in many cases these differences appear to be much less pronounced than those in the SPM dataset. In particular, there appear to be much smaller

differences between White and Non-white test takers in this dataset than in the previous one, the difference between these groups being of comparatively little practical importance. The comparisons by broad ethnic group classification show the White group to score substantially better than the Black and Middle Eastern groups, though the sample sizes within these groups are relatively small. The most striking differences observed were when groups were compared according to the reduced ethnic group classification. The White – British group scored less well than all of the other groups with the exception of the Indian group. The difference between White – British and other White test takers (again, with the exception of White – Irish test takers) was around 0.3 S.D. units, again suggesting that White-Non-white differences in the previous comparisons may be suppressed by the presence of the lower-scoring White – British test takers. The most striking difference was between White – British and Chinese test takers, the Chinese group scoring nearly 0.7 S.D. units higher than the White British majority group.

These differences were explored further in a series of linear regressions, each conceptualisation of ethnicity being investigated for its ability to satisfactorily predict test score differences. The results of these analyses are shown in Tables 27, 28 and 29:

	B	**S.E. B**	***β***
Constant	13.17	0.25	
Ethnicity	0.06	0.28	0.01

Table 27. Linear regression of ethnicity reclassified as White-Non-white predicting total APM raw score. Note: R^2 = .00

	B	**S.E. B**	***β***
Constant	13.23	0.13	
Asian	0.24	0.31	0.02
Black	-1.90	1.04	-0.05
Middle Eastern	-2.62	1.04	-0.07*

Table 28. Linear regression of broad ethnic group predicting total APM raw score. Note: R^2 = .01; *p < .05

	B	S.E. B	β
Constant	13.87	0.17	
White – British	-1.25	0.25	-0.14**
Indian	1.55	0.51	-0.11**
Chinese	1.55	0.49	0.09**
Other Asian	-0.46	0.62	-0.02

Table 29. Linear regression of reduced ethnic group classification predicting total APM raw score. Note: R^2 = .04; **p < .001

The variance explained by ethnicity in the first two analyses was minimal. However, that explained by ethnicity in the third regression model was around 4%, a comparable figure to that within the SPM dataset. There are two implications of this. Firstly, it would appear to support H_1, that there are real, measureable differences between mean test performance scores of ethnic groups within the dataset. Secondly, these results are consistent with the findings of the previous study that the conflation of ethnic groups to single classifications (such as White – British and non-British White test takers or Chinese and other Asian test takers) may be counter-productive to the understanding of ethnic group differences in test performance.

3.3.2.2 Traditional Ethnic DTF Identification using the LR Method

The item-response data in the archive was then investigated for the presence of ethnic DIF. Based on the findings in the previous analysis that ethnic differences were much more pronounced between ethnic subgroups in the dataset than they were between broad ethnic groups, it is unfortunate that the only groups with sample sizes large enough to allow ethnic DIF detection to be conducted using the LR Method.

Nevertheless, ethnic DIF between White and Non-white groups was investigated for the APM. As for the SPM archive, the data were cleansed prior to analysis to eliminate any case for which ethnicity was not recorded. This left 1507 cases in the cleansed dataset.

APM Iteration 1

	Nagelkerke R^2 values at each step in the hierarchical regression				DIF R^2 (ΔR^2	
	Step 1	Step 2 (uniform DIF)	Step 3 (non-uniform DIF)	DIF χ^2 (2) test	between Steps 1 & 3)	DIF detected
Item 1	.230	.230	.232	0.84	.002	Negligible
Item 2	.331	.331	.332	0.25	.001	Negligible
Item 3	.330	.336	.336	6.73*	.006	Negligible
Item 4	.328	.332	.333	5.65	.005	Negligible
Item 5	.205	.206	.209	4.54	.004	Negligible
Item 6	.214	.220	.220	7.40*	.006	Negligible
Item 7	.173	.173	.174	0.92	.001	Negligible
Item 8	.327	.328	.328	1.40	.001	Negligible
Item 9	.361	.366	.366	8.00*	.005	Negligible
Item 10	.370	.372	.373	4.40	.003	Negligible
Item 11	.237	.238	.239	3.22	.002	Negligible
Item 12	.254	.256	.257	4.55	.003	Negligible
Item 13	.223	.225	.226	3.84	.003	Negligible
Item 14	.279	.295	.300	29.51***	.031	Negligible
Item 15	.248	.250	.251	3.75	.003	Negligible
Item 16	.269	.269	.270	0.26	.001	Negligible
Item 17	.338	.338	.338	0.34	.000	Negligible
Item 18	.333	.333	.334	0.67	.001	Negligible
Item 19	.341	.341	.341	0.19	.001	Negligible
Item 20	.244	.245	.245	0.91	.001	Negligible
Item 21	.348	.349	.349	1.56	.001	Negligible
Item 22	.276	.277	.277	1.25	.001	Negligible
Item 23	.321	.324	.324	2.39	.003	Negligible

Table 30. Initial iteration for LR Method identification of White-Non-white ethnic DIF in the APM. Note: *$p < .05$; **$p < .01$; ***$p < .001$

No significant ethnic DIF was detected in the first iteration. The analysis, therefore, terminated after this iteration. As in the analyses for the SPM, it was again deemed necessary to examine the combined effects of these DIF effects at the test level. TCCs were generated for focal and referent groups. These are show in Figure 18 below:

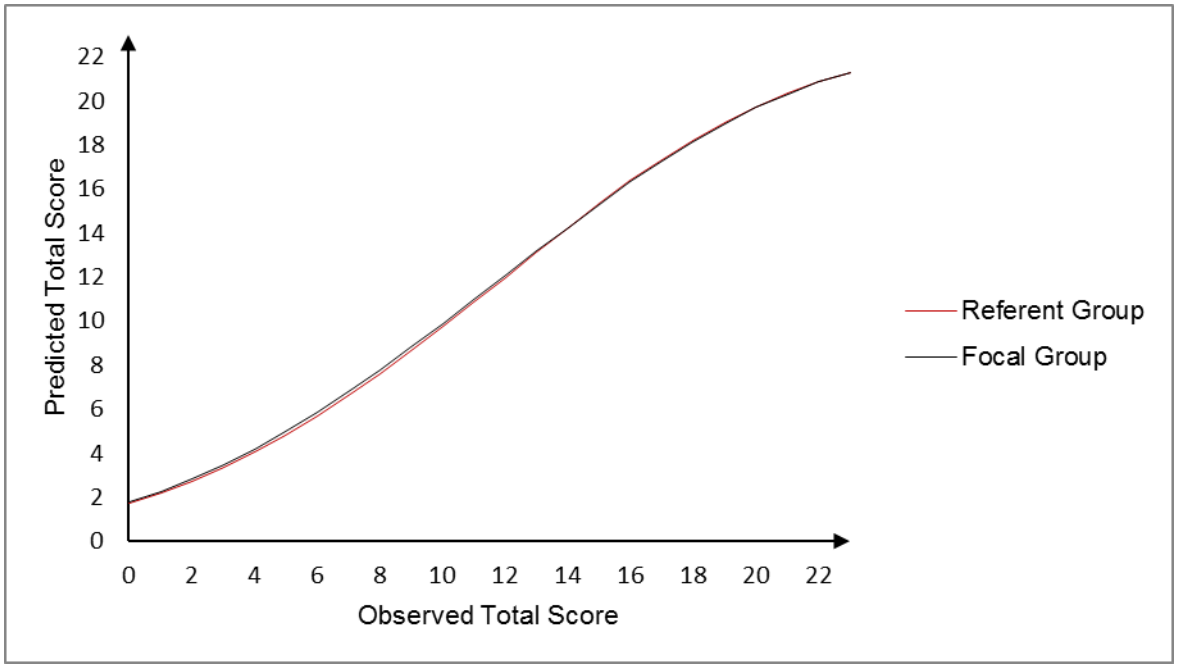

Figure 18. Test Characteristic Curves based on LR DIF parameters for White (Referent) and Non-white (Focal) groups responding to items of the APM.

Examining Figure 18, there is very little difference between the TCCs for focal and referent groups, a mean difference of only 0.05 of a raw score point in favour of the Non-white group existing across all points of the curve. The largest difference between the curves is at an observed score of 8 raw score points, but this only represents a difference of 0.17 of a raw score point, again favouring the Non-white group. Given that the mean test performance difference between these groups was 0.06 of a raw score point in favour of the White group, this DTF effect is unable to explain these performance differences.

The only other ethnic group that was large enough to compare against the White group was the Asian group (N = 238).

APM White-Asian DIF Analysis – Iteration 1

	Nagelkerke R^2 values at each step in the hierarchical regression					
	Step 1	Step 2 (uniform DIF)	Step 3 (non-uniform DIF)	DIF χ^2 (2) test	DIF R^2 (ΔR^2 between Steps 1 & 3)	DIF detected
Item 1	.220	.221	.223	5.17	.003	Negligible
Item 2	.310	.310	.310	0.21	.000	Negligible
Item 3	.316	.324	.325	7.95*	.009	Negligible
Item 4	.323	.331	.332	8.31*	.009	Negligible
Item 5	.204	.204	.207	4.13	.003	Negligible
Item 6	.211	.211	.211	11.10**	.000	Negligible
Item 7	.172	.172	.173	1.02	.001	Negligible
Item 8	.317	.318	.318	0.92	.001	Negligible
Item 9	.358	.362	.362	5.10	.004	Negligible
Item 10	.373	.377	.379	8.48**	.006	Negligible
Item 11	.229	.232	.232	4.41	.003	Negligible
Item 12	.251	.253	.255	5.48	.004	Negligible
Item 13	.215	.222	.222	8.57*	.007	Negligible
Item 14	.274	.288	.294	27.48***	.020	Negligible
Item 15	.242	.245	.246	4.60	.004	Negligible
Item 16	.268	.270	.270	1.95	.002	Negligible
Item 17	.333	.334	.334	1.68	.001	Negligible
Item 18	.327	.327	.327	0.71	.000	Negligible
Item 19	.331	.331	.331	0.68	.000	Negligible
Item 20	.237	.238	.239	1.93	.002	Negligible
Item 21	.351	.353	.354	0.05	.003	Negligible
Item 22	.269	.272	.272	3.77	.003	Negligible
Item 23	.332	.333	.335	2.70	.003	Negligible

Table 31. Initial iteration for LR Method identification of White-Asian ethnic DIF in the APM.
Note: *$p < .05$; **$p < .01$; ***$p < .001$

Though some significant ethnic DIF effects were observed, all of these were negligible in terms of their effect size. The analysis terminated after this iteration. TCCs were then generated for focal and referent groups. These are shown in Figure 19 below:

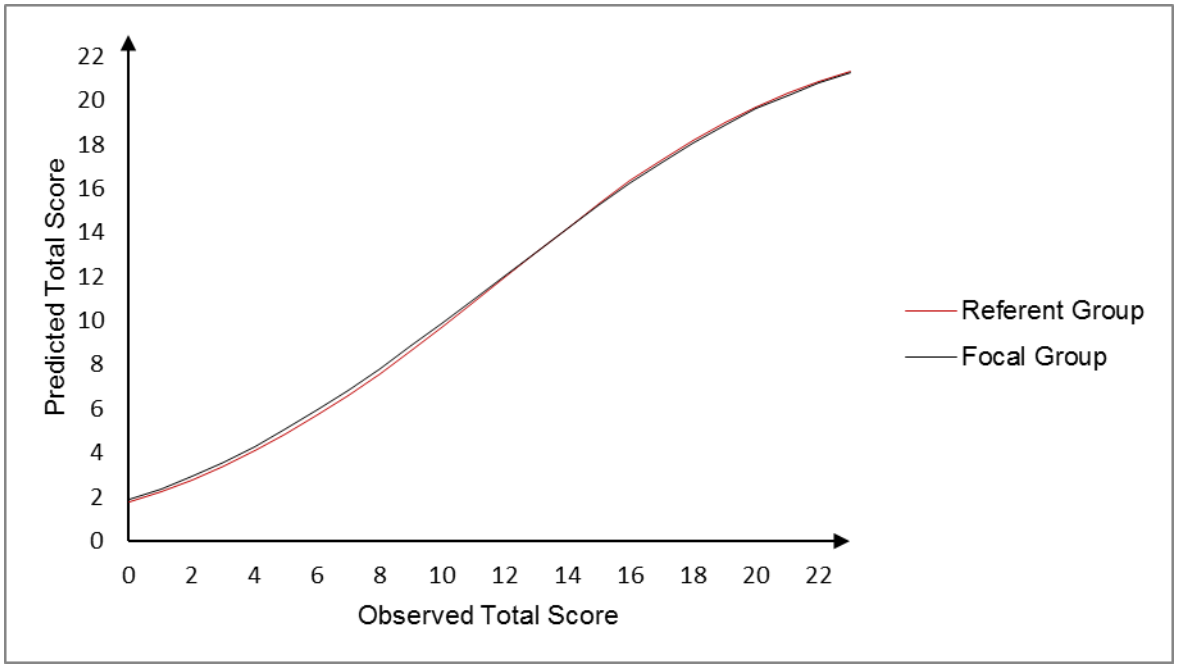

Figure 19. Test Characteristic Curves based on LR DIF parameters for White (Referent) and Asian (Focal) groups responding to items of the APM.

Examining the TCCs in Figure 19, there is very little difference in the form of either curve. The mean difference between the focal and referent group equates to 0.06 of a raw score point in favour of the Asian group across all observed scores, the largest difference between them being 0.24 of a raw score point in favour of the Asian group at an observed score of 6 raw score points. This difference is the same as the mean test performance between these groups observed in section 3.3.2.1. However, this does not provide particularly compelling evidence that ethnic DTF can account for these group performance differences, given the very small difference in test performance between these groups.

On the basis of all of these analyses, it would appear, again, that meaningful DTF could not be detected due to the limitations of the LR Method.

3.3.2.3 DTF Analysis using MLVM

Thankfully, the nature of DTF analysis using the MLVM procedure does not suffer from the same limitations. As before, a MLVM was run on the item response data in the archive and the fit statistics for one-, two-, three-, and four-class models examined to identify the most likely number of distinct response patterns that existed within the dataset. Prior to this, the assumption of unidimensionality was again assessed using EFA with unweighted least squares estimation and PROMAX rotation, based on tetrachoric correlations. The eigenvalue of the first factor extracted was 7.30. However, as was the case for the SPM, the Kaiser criterion indicated a 5-factor structure underlying the APM based on the number of factors extracted with eigenvalues greater than 1. The scree plot of eigenvalues was generated, and is shown in Figure 20 below:

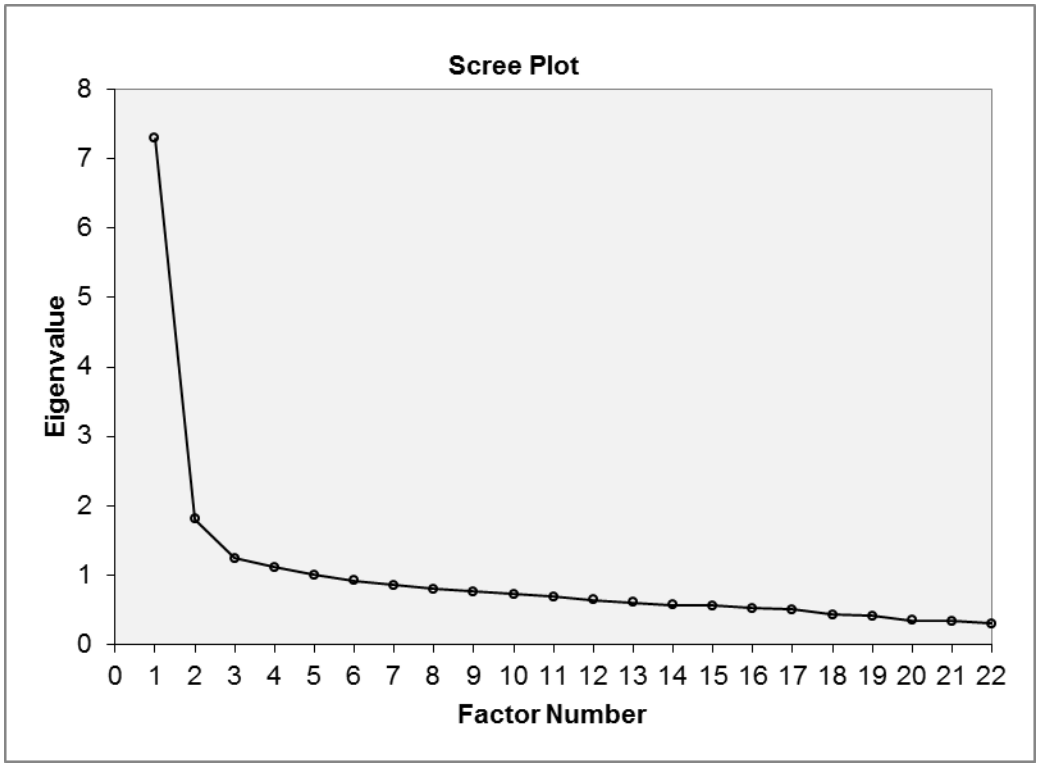

Figure 20. Screen Plot of eigenvalues for factors underlying the APM.

Examining Figure 20, the plot shows a clear point of inflexion at the 3rd factor. As was the case for the SPM, this indicates that two factors underlie the APM. To confirm this, PA-MRFA was conducted on the data. The analysis pointed to two factors underlying the data (with eigenvalues of 7.30 and 1.81), jointly explaining 43.9% of the variance. This, again, potentially represents a breach in the assumptions for MLVM, though careful analysis of the obtained factor solution was needed to confirm whether this second factor represents an additional substantive latent factor.

In spite of this, the analysis was continued to investigate whether evidence of sample heterogeneity could be found. The fit statistics generated for the 1-, 2-, 3- and 4-class models are shown in Table 32 below:

	AIC	**BIC**	**Sample-adjusted BIC**	**Entropy**
One class	36588.77	**36833.39**	36687.26	-
Two-class	36364.02	36858.58	36858.58	.52
Three-class	**36313.71**	37058.21	**36613.47**	.60
Four-class		Unidentified		

Table 32. Fit statistics generated by Mplus for the 1-, 2-, 3- and 4-class models.

In contrast to the SPM dataset, class enumeration was much clearer based on the fit statistics generated. Of the four models, the 3-class model appeared to be the best fit to the data, showing the lowest value for the AIC and sample-adjusted BIC. The 3-class model's entropy value of .60 shows improvement over the 2-class model in its ability to categorise participants, indicating only moderate overlap between classes. The 2-class model shows no improvement in BIC and sample-adjusted BIC over the 1-class model, and a poorer AIC, sample-adjusted BIC and entropy value compared to the 3-class model, so can likely be discounted. As for the 3-class model in the SPM dataset, the 4-class model was unidentified.

The model parameters generated for the each latent class in the 3-class solution were then compared to those of the one-class solution. These are shown in Table 33 below:

Item	One-class			Class 1			Class 2			Class 3		
	λ	τ	b	λ	τ	b	λ	τ	b	λ	τ	b
1	1.34	-3.77	-2.81	1.30	-3.26	-2.50	0.26	-4.17	-16.04	241.66	-491.72	-2.04
2	1.82	-4.18	-2.30	2.10	-3.88	-1.85	0.03	-4.21	-145.74	0.15	-4.58	-31.18
3	1.50	-2.45	-1.63	1.90	-2.06	-1.08	0.47	-3.23	-6.86	-0.55	-2.34	4.22
4	1.47	-2.28	-1.55	1.29	-1.68	-1.31	2.50	-4.54	-1.81	0.52	-2.02	-3.88
5	0.82	-1.14	-1.39	0.99	-1.09	-1.11	0.75	-1.32	-1.76	-0.20	-0.88	4.39
6	0.85	-1.13	-1.34	0.73	-0.72	-0.99	0.89	-1.76	-1.98	0.93	-1.68	-1.79
7	0.70	-1.07	-1.54	0.55	-0.55	-1.00	1.32	-2.16	-1.63	-1.55	-2.73	1.76
8	1.26	-1.14	-0.90	1.34	-0.59	-0.44	0.86	-2.35	-2.74	0.66	-0.44	-0.66
9	1.36	-0.88	-0.65	1.42	-0.39	-0.28	1.11	-1.55	-1.40	0.90	-1.13	-1.26
10	1.45	-1.17	-0.81	1.11	-0.51	-0.46	2.11	-2.95	-1.40	4.19	-2.32	-0.55
11	0.86	-0.01	-0.01	0.98	0.39	0.40	0.50	-0.68	-1.37	0.50	0.36	0.71
12	0.92	-0.45	-0.49	0.84	-0.04	-0.05	0.86	-1.15	-1.34	0.57	-0.32	-0.57
13	0.80	-0.21	-0.26	0.82	-0.10	-0.13	0.49	-0.79	-1.63	5.05	4.88	0.97
14	0.98	0.21	0.22	1.23	0.15	0.12	0.61	-0.14	-0.23	10.66	15.83	1.49
15	0.90	0.71	0.79	0.87	0.76	0.87	0.88	0.38	0.44	0.57	1.77	3.12
16	0.95	0.67	0.71	1.22	0.52	0.43	0.78	0.52	0.67	-1.41	4.34	-3.08
17	1.20	0.06	0.05	1.01	0.06	0.06	1.87	0.22	0.12	1.85	-0.40	-0.21
18	1.16	-0.01	-0.01	1.20	-0.25	-0.21	1.40	0.13	0.09	1.19	0.98	0.82
19	1.24	0.59	0.47	0.95	0.53	0.56	2.54	0.93	0.37	1.42	0.56	0.39
20	0.89	0.54	0.61	0.83	0.38	0.45	2.19	1.37	0.63	0.07	0.11	1.67
21	1.26	0.81	0.65	1.17	0.42	0.36	3.19	2.24	0.70	1.05	1.48	1.41
22	1.07	0.84	0.79	0.81	0.79	0.98	2.56	1.75	0.68	2.64	-0.23	-0.09
23	1.47	2.73	1.86	0.81	2.60	3.23	3.72	4.39	1.18	3.96	3.92	0.99

Table 33. MLVM parameters generated for the APM (3-class solution).

Examining the parameters in Table 33, it appears that there are substantial differences in the item factor loadings and thresholds across latent classes, particularly for members of Class Three. Though the magnitude of these parameters is extreme for Item 1, the rest of these parameters appear to be within a sensible range when compared to those obtained by Sawatzky et al. (2012) in their most defensible model. On the basis of these observations, the 3-class model was deemed to be defensible.

TCCs were then generated to demonstrate the implications for ignoring sample heterogeneity, according to the method described in section 3.2.3.3. These TCCs are shown in Figure 21 below:

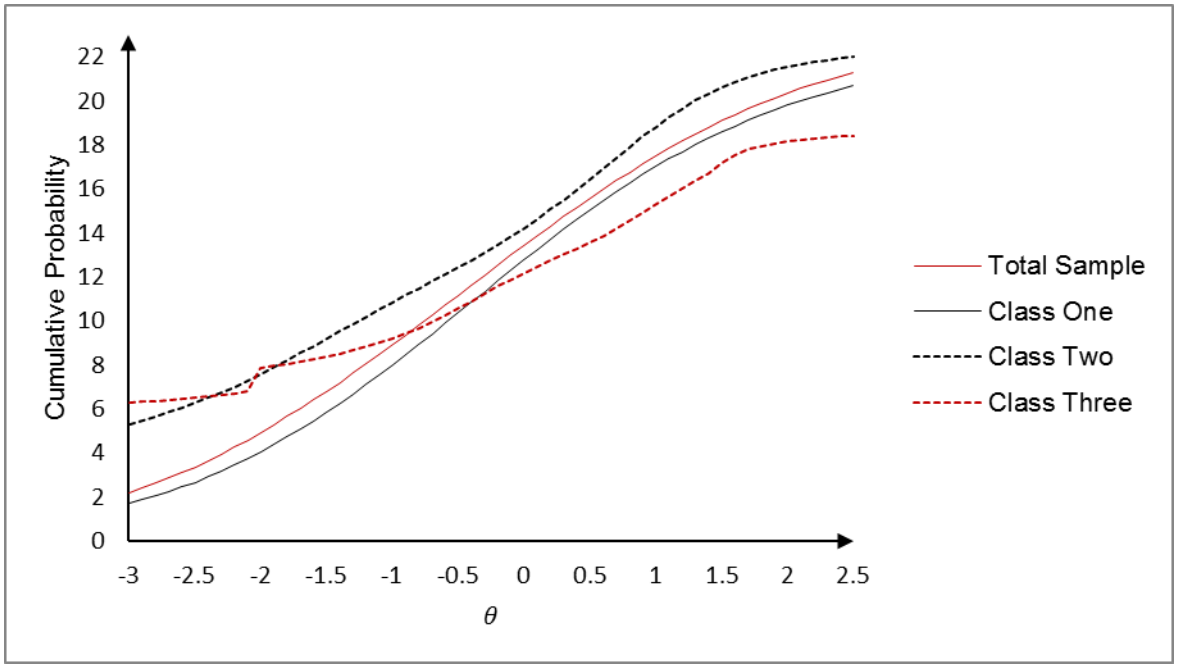

Figure 21. TCCs for each latent class in the 3-class model and for the total sample (One-class model).

Examining the TCCs in Figure 21, Class Two appear to have a clear advantage over Class One at all levels of θ. This effect is most pronounced at very low levels of θ. By contrast, the form of the TCC for Class Three appears non-uniform, in that its members appear to have an advantage over both Class One and Class Two at very low levels of θ, but are disadvantaged at moderate to high levels. All of this suggests that the heterogeneity in the sample gives rise to non-ignorable DTF effects across latent classes.

Finally, to explore the sources of class heterogeneity, the class membership variable generated by the 3-class MLVM was examined to investigate the differences in terms of size and test performance between latent classes. These results are shown in Table 34 below:

3-class Model	N	Mean APM Raw Score (S.D.)	*z*-score
Class 1	835	12.49 (4.63)	-0.15
Class 2	530	14.64 (4.05)	0.33
Class 3	142	12.19 (3.21)	-0.22

Table 34. Number, mean raw score and z-score for each class within the 3-class and 4-class models.

Examining Table 34, response patterns again appear to lead to differences in test performance that are not based on differences in cognitive ability. Though performance differences between Classes 1 and 3 in the 3-class model are fairly minimal, Class 2 outperforms both of the other two classes by around 2 raw score points.

The classes were then inspected using crosstabs to ascertain the ethnic breakdown within each of the classes. These are shown in Tables 35, 36 and 37 below:

3-class Model	White N (% of group)	Non-white N (% of group)
Class 1	666 (55.6%)	169 (54.5%)
Class 2	431 (36%)	99 (31.9%)
Class 3	100 (8.4%)	42 (13.5%)

Table 35. Frequency of White and Non-white participants within each latent class.

3-class Model	White N (% of group)	Asian N (% of group)	Black N (% of group)	Middle Eastern N (% of group)	Latin-American N (% of group)
Class 1	666 (55.6%)	124 (52.1%)	13 (72.2%)	11 (61.1%)	2 (50%)
Class 2	431 (36%)	76 (31.9%)	5 (27.8%)	6 (33.3%)	2 (50%)
Class 3	100 (8.4%)	38 (16%)	0 (0%)	1 (5.6%)	0 (0%)

Table 36. Distribution of broad ethnic groups within each class.

3-class Model	White British N (% of group)	Other White N (% of group)	Indian N (% of group)	Chinese N (% of group)	Other Asian N (% of group)
Class 1	318 (56.9%)	331 (54%)	46 (57.5%)	41 (46.6%)	25 (48.1%)
Class 2	189 (33.8%)	237 (38.7%)	22 (27.5%)	35 (39.8%)	15 (28.8%)
Class 3	52 (9.3%)	45 (7.3%)	12 (15%)	12 (13.6%)	12 (23.1%)

Table 37. Distribution of reduced ethnic categories within each class.

These tables again show that there is little difference in the distribution across classes between the ethnic groups in the dataset. It would appear, then, that ethnicity is once again independent of response behaviour.

This was investigated further in a series of logistic regressions. The procedure for this is similar to that employed when investigating the 2-class model for the SPM. However, in this

case, as the class membership variable contains more than two categories (i.e. 3), the variables were instead entered into a multinomial logistic regression model. In each model, the class membership variable generated by Mplus for the 3-class model was regressed on to ethnicity. The outputs of the logistic regressions are shown in Tables 38, 39 and 40 below:

		95% CI for Odds Ratio		
	B (SE)	***Lower***	***Odds Ratio***	***Upper***
Class 2 vs. Class 1				
Intercept	-0.44 (0.06)			
Ethnicity	-0.10 (0.14)	0.69	0.91	1.19
Class 3 vs. Class 1				
Intercept	-1.90			
Ethnicity	0.50	1.11	1.66	2.46

Table 38. Multinomial logistic regression for White-Non-white classification predicting latent class in the 3-class model. Note: R^2 = .01 (Cox & Snell), .01 (Nagelkerke). Model $\chi^2(2)$ = 7.68, $p < .05$.

		95% CI for Odds Ratio		
	B (SE)	***Lower***	***Odds Ratio***	***Upper***
Class 2 vs. Class 1				
Intercept	-0.61 (0.51)			
White	0.17 (0.51)	0.44	1.19	3.23
Asian	0.12 (0.53)	0.40	1.12	3.16
Middle Eastern	(Reference)			
Class 3 vs. Class 1				
Intercept	-2.40 (1.04)			
White	0.50 (1.05)	0.63	1.65	12.93
Asian	1.22 (1.06)	0.25	3.37	26.96
Middle Eastern	(Reference)			

Table 39. Multinomial logistic regression for broad ethnic classification (Latin-American and Black excluded) predicting latent class in the 3-class model. Note: R^2 = .01 (Cox & Snell), .01 (Nagelkerke). Model χ^2 (4) = 12.30, $p < .05$

	B (SE)	*95% CI for Odds Ratio*		
		Lower	***Odds Ratio***	***Upper***
Class 2 vs. Class 1				
Intercept	-0.51 (0.33)			
White – British	-0.01 (0.34)	0.51	0.99	1.93
Other White	0.18 (0.34)	0.62	1.19	2.31
Indian	-0.23 (0.42)	0.35	0.80	1.81
Chinese	0.35 (0.40)	0.65	1.42	3.11
Other Asian	(Reference)			
Class 3 vs. Class 1				
Intercept	-0.73 (0.35)			
White – British	-1.08 (0.38)	0.16	0.34	0.72
Other White	-1.26 (0.39)	0.13	0.28	0.60
Indian	-0.61 (0.48)	0.21	0.54	1.39
Chinese	-0.50 (0.48)	0.24	0.61	1.56
Other Asian	(Reference)			

Table 40. Multinomial logistic regression for reduced ethnic classification predicting latent class in the 3-class model. Note: R^2 = .02 (Cox & Snell), .02 (Nagelkerke). Model χ^2 (8) = 21.02, p < .01

Though all three of the regression models in this set of analyses were significant, the pseudovariance explained by each was very small. The implication of these results is that, as has been observed for the SPM, the APM does not display ethnic DTF.

In an attempt to explain what these latent classes represent, the gender, age, and educational level were entered into multinomial logistic regression models predicting class membership. As for the SPM, none of these variables were able to explain more than trivial pseudovariance in class membership.

Odds ratios for each item were once again generated for each latent class, based on the thresholds obtained for the 3-class MLVM. This made interpretation of the pattern of odds ratios somewhat more complex than it was for the SPM, as, for each item, three odds ratios were generated, based on comparisons of the odds of a correct response between each pair of classes. The odds ratios generated for each item are shown in Table 41.

	Odds Ratios		
Item	Class 1 vs. Class 2	Class 1 vs. Class 3	Class 2 vs. Class 3
1	0.40	-	-
2	0.72	0.50	0.69
3	0.31	0.76	2.44
4	0.06	0.71	12.43
5	0.79	1.23	1.55
6	0.35	0.38	1.08
7	0.20	0.11	0.57
8	0.17	1.16	6.75
9	0.31	0.48	1.52
10	0.09	0.16	1.88
11	0.34	0.97	2.83
12	0.33	0.76	2.29
13	0.50	145	290
14	0.75	6452640	8623486
15	0.68	2.75	4.01
16	1.00	45.60	45.60
17	1.17	0.63	0.54
18	1.46	3.42	2.34
19	1.49	1.03	0.69
20	2.69	0.76	0.28
21	6.17	2.89	0.47
22	2.61	0.36	0.14
23	5.99	3.74	0.63

Table 41. Odds ratios comparing odds for each latent class of a correct response to each item in the APM. Note: Odds ratios greater than 1 indicate higher odds of a correct response for the first class in each comparison.

Examining Table 41, the odds ratios of correct responses between pairs of latent classes is somewhat more challenging. As for the SPM, the majority of odds ratios represent effects that are of small or trivial magnitude, though there are ten items for which one or more medium or large effects were observed, namely Items 4, 7, 8, 10, 13, 14, 16, 20, 21, 22 and 23. For example, for Items 13, 14 and 16, members of both Class One and Class Two have much higher odds of correctly responding to the item than do members of Class Three. For Item 1, the odds ratios between Class 3 and the other two classes are undefined, as no members of Class 3 responded incorrectly to the item. This meant that calculation of odds of a correct response to this item for members of Class Three involved division by zero. However, it can be inferred without calculation that Class Three has the highest odds of

responding correctly to Item 1 relative to the other classes, and that both of these differences would represent large effects.

It is, however, easier to appreciate differences in response behaviour across classes by considering effects across all items, rather than effects in individual instances. Once again, across all items in the APM, a general pattern emerges. If the odds ratios for each pair of classes are used to rank order the likelihood of each class' odds of correctly responding to each of the items, it appears that, for the vast majority of early items in the APM, members of Class Two have the highest odds of a correct response. By comparison, in these earlier items, members of Class One have the lowest odds of a correct response. However, this pattern appears to reverse for the later items, where it can be observed that members of Class One tend to have the highest odds of a correct response, and Class Two the lowest. Conversely, the odds of members of Class Three responding correctly relative to members of the other two classes appears to vary throughout the test, showing no such consistent pattern of advantage or disadvantage relative to those of the other two classes. It would appear, then, that there are two distinct properties within the items of the APM. One of these properties is more prevalent in the earlier items and leads to members of Class Two performing better than their peers. The other is more prevalent in the later items and appears to favour member of Class One. The nature of these properties is explored more fully in section 3.4.2.

3.3.2.4 MLVM on Unifactorial Item Subsets within the APM

As was the case for the SPM, the EFA conducted on the items of the APM suggested that this tool might not be, in actual fact, unidimensional. It is possible, then, that the evidence uncovered for sample heterogeneity in the main analyses may have been due to model misspecification, as appeared to be the case for the SPM.

To examine this more thoroughly, subtests were constructed from the items that loaded most strongly on to the two factors underlying the APM. To do this, the PROMAX-rotated factor matrix generated by the EFA was examined.

	Factor	
	1	**2**
Item 1	**0.74**	-0.20
Item 2	**0.66**	
Item 3	**0.77**	
Item 4	**0.71**	
Item 5	**0.41**	
Item 6	**0.45**	
Item 7	**0.44**	
Item 8	**0.64**	
Item 9	**0.52**	
Item 10	**0.48**	0.20
Item 11	**0.49**	
Item 12	**0.34**	
Item 13	**0.33**	
Item 14	**0.30**	0.25
Item 15	0.22	**0.30**
Item 16	**0.30**	0.26
Item 17		**0.51**
Item 18		**0.45**
Item 19		**0.65**
Item 20		**0.63**
Item 21		**0.70**
Item 22		**0.62**
Item 23		**0.85**

Table 42. Rotated factor matrix for the 2-factor solution underlying the APM. Note: Factor loadings below .20 have been suppressed.

On the basis of these factor loadings, two subtests were constructed from the APM's items. These were then reanalysed to examine whether DTF could be identified for these subtests.

Factor 1

The Factor 1 subtest for the APM comprised 15 items. These items were Items 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, and 16. A MLVM model was conducted on this subtest, the fit statistics generated by which are shown in Table 43 below:

	AIC	BIC	Sample-adjusted BIC	Entropy
One class	23035.27	**23194.80**	**23099.50**	-
Two-class	23000.40	23324.79	23131.01	.70
Three-class	**22985.55**	23474.79	23182.54	.73
Four-class	22986.69	23640.79	23250.05	.56

Table 43. Fit statistics generated using a reduced set of APM items based on loadings on the first extracted factor.

Examining these fit statistics, the same pattern appears as in the analyses for the SPM subtests. The AIC was lowest for the three-class model, but this was offset by substantially lower BIC and sample-adjust BIC values for the one-class model. This suggests that there is no DTF within the Factor 1 subtest.

Factor 2

Finally, the Factor 2 subtest was examined to see if the sample heterogeneity observed in the main analyses could be attributed to this subset of items. There were 7 items within this subtest: Items 15, 17, 18, 19, 20, 21, 22, and 23. A MLVM was conducted on this subtest, the fit statistics generated for which are shown in Table 44.

	AIC	BIC	Sample-adjusted BIC	Entropy
One class	13910.20	**13995.28**	**13944.45**	-
Two-class	**13877.18**	14052.67	13947.84	.23
Three-class	13881.17	14147.06	13988.23	.73
Four-class	13886.83	14243.13	14030.29	.62

Table 44. Fit statistics generated using a reduced set of APM items based on loadings on the second extracted factor.

The fit statistics and entropy values shown in the table above suggest, once again, that no improvement in model fit can be gained by fitting a more complex model to the data than the one-class model. This suggests, overall, that there is no DTF present in the APM, and that the effects observed in the main analyses are, most likely, spurious.

3.4 Discussion

3.4.1 Ethnic Differences in Test Performance

Examining the descriptive statistics by ethnic group of test performance, and the variance that can be explained in total test score by ethnicity, one thing is clear: Persistent differences exist between ethnic groups in terms of their mean test performance. Furthermore, ethnicity was found to explain a small-yet-meaningful proportion of the variance in test scores for both the SPM and the APM. A key finding from the present study was that, in many cases, these ethnic group differences are substantially smaller than those that have previously been observed in the literature. At 0.9 S.D. units, the most extreme of the differences observed – that between the White majority group and the Black group in the SPM dataset – was less than the classically-observed difference of around 1 S.D. (e.g. Gottfredson, 2005; Rushton & Jensen, 2005), and this difference was much less pronounced for the APM at less than 0.5 S.D. units. More broadly, in the majority of cases, ethnic group test scores fall short of the expected 0.46 – 0.83 S.D. units difference with the White majority, based on the literature (Martocchio & Whitener, 1992; Schmitt, Clause & Pulakos, 1996).

One possible explanation for this disconnect between differences observed in the literature and those found in the present study is the gradual narrowing of group test score differences over time (Dickens & Flynn, 2006). The vast majority of research conducted on ethnic group test score differences was published ten or more years ago, often being based on data that are even older. The findings of the present study may represent further evidence of the narrowing of ethnic group difference gaps over time, based on more recent data.

While these particular findings are undoubtedly encouraging, they should be tempered by consideration of the practical impact that these persistent differences are likely still to have in real selection processes. These relatively small differences could still represent adverse impact on certain ethnic groups, particularly if deployed in European countries outside of Britain. Differences between the Non-British White group and Indian, Pakistani and Other Asian groups all amounted to around 20 percentile points in the SPM. While the actual impact of this depends heavily on at which point the cut score for a test in a particular selection process is set, even a conservative cut score might risk adverse impact on one or all of these groups if it were to produce large enough differences in selection ratios (e.g. Bobko & Roth, 2009).

Comparing the two tests, differences in terms of raw score points were much larger for the SPM than for the APM. This might, on first inspection, appear to be a function of the nature of the tests themselves: The APM is much more difficult than the SPM and contains fewer items, leading to lower overall scores and smaller raw score point differences between groups. However, ethnic differences between the highest- and lowest-scoring ethnic subgroups were of comparable magnitude when judged by their values of Cohen's *d*. It seems, then, that inconsistencies between the findings for the SPM and APM datasets in terms of comparative performance between ethnic groups has less to do with the test itself and perhaps rather more to do with the ethnic make-up of each dataset and how the conceptualisations of ethnicity impacted upon them.

In the SPM dataset, the conceptualisation of ethnicity used in the analyses seemed not to influence the explained variance in raw scores very much, ranging from around 2% to around 3.5%. Conversely, analyses of the APM dataset showed much more sensitivity to the conceptualisation of ethnicity used to examine explained variance in test scores. When ethnicity was represented as White-Non-white and by broad ethnic group, the explained variance in test scores was less than 1%. However, when these broad ethnic groups were broken down into the subgroups they represented (or, rather, those of a sufficiently large sample size to meet the study's inclusion criteria), the variance explained in total raw score was nearly 4%.

Perhaps the most striking difference between ethnic groups – and one that was consistent across both archives – was that between the British subgroup of the White ethnic group and all other subgroups within it. White British test takers were outperformed by Non-British White test takers on both the SPM (d = .35) and the APM (d = .29) by around a third of an S.D. unit. The explanation for why this might be the case is not immediately clear, though is likely to be socio-environmental in nature. One possibility may lie in the quality of education in different countries. It has been consistently found that pupils in the UK are lagging behind their European peers in terms of academic achievement. For example, the UK was found to be substantially outperformed by the USA, Canada, The Netherlands, Australia, Germany and Sweden on tests of middle school level mathematics and science (measured by the TIMMS), and adult literacy as measured by the IALS (Barro & Lee, 2001). While the differences observed in this study between White groups is not massive, it has an important implication for the investigation of ethnic bias: the conflation of White subgroups may serve – in datasets where White British test takers are the majority – to suppress the magnitude of differences between the highest- and lowest-scoring ethnic groups. This was not the only example of an ethnic sub-group outperforming other subgroups within their ethnic classification. In the APM dataset, Chinese test takers scored much more highly than other groups within the Asian broad ethnic classification, scoring over half an S.D. unit higher than

the White British majority, compared to a difference of only 0.18 S.D. units above the majority for Other Asian test takers. However, the difference between White and Asian ethnic groups in the APM dataset only amounted to a difference of .2 of a raw score point (*d* = 0.05 in favour of the Asian group).

The implication of this is that it may be, in some cases, counterproductive when examining ethnic differences in test performance to consider these differences at the broad group level. At the very least, before reducing ethnic groups to broad classifications, researchers would need to examine the effect that this is likely to have based on the mean group test scores of differing ethnic subgroups. If subgroups are found to perform substantially differently to one another, it makes little sense to consider them as differing parts of the same whole. Indeed, this default approach to the simplification of ethnicity through the conflation of groups on the basis of manifest physical characteristics has previously been criticised by Mason (2000). Despite this, it has been noted that this approach is still very widely used in the ethnic bias literature (Kenny & Briner, 2007). Instead, if the higher- and lower-scoring subgroups are sufficiently large for the purposes of the study, it might be preferable to treat them as separate ethnic groups. If, however, a higher- or lower-scoring subgroup represents a very small proportion of the total sample, it might be more prudent to treat members of this subgroup as outliers within the broad ethnic group to which they belong, excluding them from further analyses.

3.4.2 Ethnic DIF and DTF in the Raven's Measures

Having established that ethnic group test score differences existed, the study moved on to consider what proportion of the variance in test performance explained by ethnicity could be attributable to DTF. To do this, two competing methods for the identification of DIF and DTF were employed. The first was the LR Method of DTF identification, a more traditional method that demonstrates many advantages over other methods such as the MH Procedure

and the SIBTEST. The second, MLVM, was a more sophisticated method that examines DTF across all items of the test simultaneously, based on heterogeneity of the measurement model underlying an ability test between latent classes (i.e. differences in factor loadings and thresholds between items and the latent factor they measure between groups of candidates that are not defined *a priori*).

In both datasets, the LR Method failed to identify evidence of meaningful ethnic DTF, based on very small differences between focal and referent TCCs. It had been predicted in H_2 that this would be the case, as it has been frequently observed that traditional methods based on response differences between manifest groups (based on dichotomies such as White-Non-white, White-Black and so on) often do not detect ethnic differences in response probabilities due to the underlying assumption of these approaches that these groups are homogeneous in their response patterns (Cohen & Bolt, 2005). There is, however, a possible confounding factor that is responsible for these results. The LR Method depends upon having sample sizes of at least 200 participants in both the focal (i.e. minority) and referent (i.e. majority) groups (Zumbo, 1999). The restriction that this assumption imposed meant that only very limited ethnic dichotomies could be examined for the presence of DIF/DTF. Other than the analysis to detect White-Non-white DIF, the only other ethnic groups suitable for analyses of this type were the Asian group and, in the case of the SPM dataset, the Middle Eastern group (though, in actuality, the size of this group was slightly less than the 200 cases required). In the case of the APM analyses, the differences between these groups in S.D. units were not really large enough to yield non-null results. However, there were larger group differences in the SPM, suggesting that these results are unlikely to be due to a lack of statistical power. Nevertheless, it would have been interesting to have been able to examine ethnic DIF/DTF between more of the potential dichotomies in the dataset, particularly between Black and White test takers, given the relative large mean raw score differences between these groups. A DIF/DTF analysis between White British test takers and Other White test takers might have shed additional light on the nature of the observed

test performance differences between these groups, but without additional robust analyses of ethnic DIF between other groups, the results of this might have been difficult to contextualise.

The second technique, MLVM, also was not able to detect the presence of ethnic DTF in either dataset. The fit statistics and entropy values generated from the analyses suggest that heterogeneity exists within the datasets, implying that factor loadings and thresholds are not stable across latent classes within the sample (i.e. that, for at least some items in the tests, the probability of a participant responding correctly is not consistent across latent classes for a given level of θ). When ethnicity was used to predict latent class membership, though, in all cases the pseudovariance explained by ethnicity was found to be trivial, suggesting that ethnicity could not predict class membership with any kind of consistency. Alongside this, cross-tabulations of ethnic group by latent class revealed very little variation in the distribution of ethnic groups between latent classes. This suggests that ethnicity and latent class are largely orthogonal constructs and that, even if some form of DTF exists within the datasets (for whatever reason), it is unlikely to be able to account for ethnic group performance differences.

However, as for the LR Method, a number of methodological problems presented themselves that might well have confounded these results. The first of these to be discussed is in class enumeration, the process by which a researcher decides on the 'correct' number of latent classes that underlie the dataset. Though a decision was reached on the likely number of latent classes within both datasets that was based on empirical support from the literature, the differences in AIC, BIC and sample-adjusted BIC were very small. Furthermore, the fit statistics all pointed to a different 'correct' number of latent classes. Referring to previous attempts to use MLVM in the literature – not solely for the purposes of ethnic DTF identification – this seems to be more common than one might expect. Both Cohen and Bolt (2005) and Sawatzky et al. (2012) report fit statistics that show similar small differences between models, with only marginally better agreement between

different types of statistics. All this suggests that there is something of an art to class enumeration, that it is rarely as clear cut as perhaps it should be. Sawatzky et al. (2012) recommend that class enumeration be conducted in context, in that improvement in model fit should be interpreted in the context of how entropy values improve incrementally across models. Still, the apparent lack of clarity appears to be a limitation, as inaccurately estimating the 'true' number of latent classes within a dataset has potentially serious ramifications for how we understand the processes that underlie response behaviour.

The second limitation to the MLVM approach for the identification of DTF – specifically when this kind of analysis is conducted in Mplus – has its roots in the mathematical processes on which it is based. While the author does not claim to be a mathematician so is not qualified to critique the particulars of the algorithms underlying MLVM as a technique in itself, there are a number of potential issues pertaining to its use for DTF identification that can be identified, based on arguments from the extant literature. Its co-option for use in IRT settings – particularly those that focus on ability tests (e.g. Samuelsen, 2008) – raises some concerns, the most striking of which is in its assumption that a single latent factor is being measured by the tests used.

When this assumption was checked, EFA revealed that both the SPM and the APM most likely have a multifactorial latent structure, parallel analysis indicating that two factors underlie each test. The Raven's tests are touted as being measures of abstract reasoning ability. Furthermore, it has been demonstrated that they tend to load almost exclusively on g_f, having very little g_c component (Carpenter, Just & Shell, 1990). This being the case, it is somewhat surprising that non-trivial variance can be explained in test scores on both measures by latent factors other than pure reasoning ability. The challenge, then, becomes accounting for what these secondary factors represent. Sawatzky et al. (2012) maintain that any multifactorial solution produced by EFA needs to be carefully examined to establish whether secondary factors represent additional substantive latent factors (which would

represent a violation of the assumptions of MLVM) or random error variance (which, at least theoretically, should not).

The patterns of factor loadings for the SPM (in Table 21) and the APM (in Table 42) appear to match one another very closely in their general form. Both solutions are made up by a first factor with a large eigenvalue, onto which the majority of test items load strongly. However, in both solutions, the later items all load increasingly strongly onto a second, separate factor. These patterns appear too regular to be random. Furthermore, in both instances, the two factors are strongly correlated (.60 for the SPM; .64 for the APM). This suggests that there is a relevant additional substantive latent factor underlying the SPM and the APM, one that is related to reasoning ability, but not it in and of itself.

There are several candidates for what this mystery additional factor might represent. Carpenter, Just and Shell (1990) have previously suggested that very difficult items in the Raven's measures might tap into working memory capacity in addition to g_f. They proposed that the mechanism by which this comes about is based upon the specific nature of some of the items. Using an earlier iteration of the Raven's, they classified its items according to their difficulty, rule type (i.e. nature of the problem and solution), and the number of 'rule tokens' that each had, a rule token representing a distinct form of reasoning operation that needed to be performed on the item to solve it. They observed two important trends in these item characteristics. Firstly, item difficulty – the assessment of which they based on the error rate for each item – progressed in a linear fashion, each problem being more difficult than the last, almost without exception. Secondly, they noticed that the difficulty of an item was strongly related to the number of rule tokens needed to solve it. Based on these observations, they reasoned that, for more difficult items, success was contingent on a candidate's ability to manage multiple rule tokens in working memory.

It would, then, appear logical that this second factor could be the influence of WMC on test performance. Observing the factor loadings onto the second factor in Tables 21 and 42, the strength of loadings appears to progress linearly through the test. Furthermore, this parallels

the b parameters observed for the one-class models in Tables 11 and 33 somewhat closely, implying that as difficulty increases, so does the influence of the second factor. However, the literature is not entirely agreed on Carpenter, Just and Shell's (1990) observations of the nature of the Raven's tests. Unsworth and Engle (2005) examined their previous claims, using a structural equation model of loadings between items on the APM and OSPAN (a measure of working memory capacity developed by Turner & Engle, 1989) and two latent factors, one representing g_f and the other working memory. They found that factor loadings on working memory did not vary for the items in the linear way expected, observing that working memory could explain variance in both the more difficult items and the easier ones.

It is worth, then, exploring potential alternatives in attempting to characterise this second factor. One possibility can be suggested based on previous work by Lynn, Allik and Irwing (2003). Examining a previous (and, at 60 items, substantially longer) incarnation of the SPM, they identified three distinct cognitive operations that characterised the type of problem solving required to address individual items. They termed these *gestalt continuation* (a form of pattern completion that does not require any form of high-level reasoning), *verbal–analytic reasoning* (problems requiring a solution based on relatively simple addition or subtraction between elements of the matrix), and *visuospatial ability* (which require a solution based on perceptual reasoning). They observed that the items in the first quarter of the SPM (i.e. the very easy items) tended to load almost exclusively on the gestalt continuation factor, whereupon factor loadings began to shift more towards the visuospatial ability factor, and, finally, onto the verbal-analytic reasoning factor. Though each of these operations represents a distinct form of reasoning in its own right, Lynn, Allik and Irwing note that these factors are still superseded by *g* as a single factor underlying the SPM, having observed them to be highly intercorrelated. Interestingly, the presence of these same three factors has been independently observed in the APM (DeShon, Chan & Weissbein, 1995).

Though this three-factor structure is inconsistent with those observed in the present study, the relative ease with which the items of the SPM and APM could be classified based on this structure warrants further exploration. Examining the nature of the 28 items of the SPM, the vast majority are easily categorised according to Lynn, Allik and Irwing's structure. Interestingly, however, only the first two items can truly be classed as gestalt continuation problems, both being very basic pattern completion items. It is possible, in the process of developing a 28-item short form of the test, that the developers took the decision to eliminate many of the gestalt continuation items on the basis that their relative ease did not provide enough variance in scores between candidates to be particularly useful. The remainder can be classed as either verbal-analytic reasoning or visuospatial ability (or, on occasion, a combination of both). Most interestingly, though, is that these categories match the factor loadings observed in Table 21 very closely: It would appear that the earlier items in the test almost all require – to a greater or lesser extent – visuospatial ability to solve, whereas the later items are exclusively based on verbal-analytic reasoning. This makes for compelling evidence that the two factors observed in both the SPM and APM in the study are, in fact, visuospatial ability and verbal-analytic reasoning.

This observation can, perhaps, cast some light onto the true nature of the latent classes obtained through the MLVM analyses. For the SPM, a two-class solution was obtained. When the odds ratios of a correct response to each item were examined for Class One relative to Class Two, it appeared that Class One had generally better odds of responding to the earlier items, and Class Two had better odds of responding to the later items. A similar pattern was observed for the APM. Though a three-class solution was obtained from the MLVM, it was observed that Class Two outperformed Class One on the earlier items, and Class One outperformed Class Two on the later items. Class Three showed no consistent advantage or disadvantage over the other two classes. One could infer from this, then, that Class One in the SPM and Class Two in the APM are characterised by a higher level of visuospatial ability relative to the other classes, and that Class Two in the SPM and Class

One in the APM are characterised as having a higher level of verbal-analytic reasoning. Class Three – assuming it is not a spurious latent class (which might well be the case) – could potentially be characterised by a having a relative balance of these two latent traits.

Though these implications are, at this stage, conjecture, they received further support from the post hoc MLVM analyses that were conducted. In an attempt to overcome the apparent violation of MLVM's assumption of unidimensionality, a relatively novel technique was employed in a series of follow-up analyses. Splitting the items according to the factors on to which they loaded most strongly, MLVM was rerun. In all cases, none of the more complex models showed an improvement in model fit over the one-class model. This implies that, when the items that most strongly tap into each of the two factors underlying the SPM and APM are isolated, heterogeneity of response behaviour seems to disappear. Therefore, the most likely explanation for the observed DTF in the full item set is that the latent classes generated do not represent sample heterogeneity with respect to any irrelevant construct, but, rather, are the product of true differences in the levels of the two latent factors underlying these tools.

This has some potentially serious ramifications for previous research that has made use of this technique for DTF identification, but, in doing so, has not taken the confounding influence of a measure's multifactorial nature in to account. Tests of specific ability are generally thought to tap into a single construct, and it is this assumption that allowed bias researchers to co-opt MLVM by using its single latent variable to represent θ. However, evidence from the literature indicates that very few – if any – ability tests have a purely unifactorial underlying structure, due to the contribution of secondary latent factors to their structure (for example, the contribution of g_f to verbal reasoning, or WMC to performance on the Raven's measures). This casts very real doubt over the findings of previous research into DTF in ability tests that has used MLVM, as the latent classes these studies have identified could, very easily, be the result of violations of the assumptions of MLVM, meaning that they are likely spurious. If this were the case, the factor loadings and thresholds

observed to be conditional on latent class membership might, instead, only vary according to the degree to which individual items load on to both the primary factor and any secondary substantive factors.

When considered in their totality, these findings suggest that, in all probability, no DTF exists for the Raven's measures. The conclusion that would likely be reached on the basis of this is that Differential Test Functioning is not responsible for the observed ethnic group test performance differences that have previously been observed. While this finding may not generalise across all measures of cognitive ability, it is an important step towards bringing clarity of understanding to the ethnic measurement bias literature.

CHAPTER 4: STUDY 2

4.1 Study Overview

This chapter will describe Study 2 of the research. The main aim of the study was to identify the root causes of ethnic group test performance differences, and, in doing so, to better understand the mechanisms by which these causal factors affect candidates' performance. The study's theoretical model posits that it is differences in socio-economic factors that are the root cause of these differences, but that these factors affect a candidate's test score through a range of performance facilitating and debilitating factors.

The study administered five on-line tests of ability (under proctored testing conditions), a personality inventory, and a short demographic questionnaire to a sample of UK residents of working age. Testing conditions were designed to minimise the possibility of stereotype threat influencing the results, so that the effect of performance facilitating and debilitating factors – particularly those based on personality traits – could be more easily examined.

The sample was drawn from a single national population with the intention of controlling as much as possible for the influence of the national differences in test performance observed by Rindermann (1997). Additionally, the present study drew its sample from an unrestricted range of participants, allowing both employed and unemployed adults to be recruited. This allowed for a more accurate representation of the current state of ethnic test performance differences in the UK to be considered than is usually possible.

In the absence of any robust metric of test familiarity in the literature, a scale was designed to measure participants' familiarity with ability testing. The construct of test familiarity that the scale sought to measure was based on a modified conceptualisation of Reeve, Heggestad and Lievens' (2009) definition. In this new definition, test familiarity is viewed as a construct that is not specific to a single test, but one that builds with exposure to different forms of test to allow a candidate to transfer performance-facilitating skills to new tests. Items to capture this construct were generated and then checked for their psychometric robustness as a part of the whole scale. The resultant scale is a simple, short measure that

can be used to quickly capture a variable that has classically been reduced in the literature to little more than practice effects (e.g. Anastasi, 1981; Hausknecht et al., 2007).

The chapter will begin by describing the method and measures used for the study. The analyses of the study will then be reported upon, focusing first on uncovering the nature of possible performance facilitating and debilitating factors in the sample as a whole, before exploring the extent to which these factors are able to account for ethnic group test score differences.

4.1.1 Hypotheses

On the basis of the theoretical model described in section 2.5.3, a number of hypotheses were generated for the present study:

H_1: Ethnic group differences will exist across all measures of ability.

H_2: These differences will become more pronounced as a measure's g-loading increases.

As in the previous study, the extant literature overwhelmingly points to the existence of ethnic group test score differences. Furthermore, it is expected that the Spearman-Jensen Effect will be replicable in the present study, once each test's loading on to a single latent variable has been estimated in a CFA model.

H_3: Socio-economic variables will be positively correlated with all measures of ability.

H_4: Test familiarity will be positively correlated with all measures of ability.

H_5: Test familiarity will partially mediate the relationships between socio-environmental variables and test performance.

Based on previous findings in the literature, socio-economic variables are expected to be positively correlated with performance on all the measures of ability in the present study. Given the nature of the development of cognitive ability – in that the greater part of it occurs during childhood (McArdle at al., 2002) – it is predicted that familial social status will be more strongly related to ability than is current participant social status. Parallel to this, test familiarity is expected to be positively correlated to all socio-economic variables in the study through the mechanism of exposure. This mechanism is not expected to be able to explain the entirety of the variance between socio-economic variables and test performance, as it is likely that differences in socio-economic status will be predictive of differences in true ability, particularly at the facet level of participant educational background. However, test familiarity is expected to mediate a substantial proportion of this effect.

H_6: The degree to which test familiarity can explain variance in a test's mean raw scores will be proportional to the degree to which that test taps into g_f.

The theoretical model for the study proposes that test familiarity facilitates test performance through the development of schemata. Retention of these schemata allow a participant to more efficiently recognise the general forms of test items and the rules that govern them, allowing that participant to make swifter progress through a test than if they were unfamiliar with testing. If this is the mechanism by which test familiarity facilitates performance, it stands to reason that the greatest benefits to performance would be for those tests that require the greatest degree of pattern recognition (i.e. tests with a strong fluid intelligence component). In practice, this means that a much larger proportion of the variance in test scores that can be explained by ethnicity will be attributable to test familiarity for the WMC measure and for other measures known to load on g_f than for the verbal reasoning measure used in the study, which is likely to tap more into g_c.

H_7: Key personality traits will predict substantial variance in test performance.

It is predicted that the findings represented in Chamorro-Premuzic and Furnham's (2004) model of the interaction between personality traits and ability will be replicable. Specifically, the present study hypothesises the following links between personality traits and test performance:

H_{7a}: Traits that map conceptually on to Emotional Stability will be positively correlated with test performance.

H_{7b}: Traits that map conceptually on to Openness will be positively correlated with test performance.

H_{7c}: Traits that map conceptually on to Conscientiousness will be negatively correlated with test performance.

H_{7d}: Traits that map conceptually on to Agreeableness will be uncorrelated with test performance

H_{7e}: Traits that map conceptually on to Extraversion will be positively correlated with test performance.

H_{7f}: A curvilinear regression model will be a substantially better fit to the data than a linear regression model when Stability is regressed on to test performance.

H_{7g}: A curvilinear regression model will be a substantially better fit to the data than a linear regression model when Optimism is regressed on to test performance.

Based on the Trait Personality Inventory's underlying factor structure (ABA, 2011), predictions about how specific Trait scales will be associated with test performance can be made. Traits related to Extraversion (Sociability; Leadership; Achievement; Optimism) are expected to be positively correlated with test performance due to these traits' performance

facilitative effects in the form of lower arousal (leading to less susceptibility to distraction) and more confidence (leading to shorter overall item response time). The Trait scale that represents the anxiety component of Emotional Stability (Stability) is expected to be positively correlated with test performance, due to the test performance debilitative effect of test anxiety. Traits related to Openness (Intellect; Culture) are expected to be positively related to test performance, but only for those measures that have a strong g_c component (expected to be the measure of verbal reasoning used in the study). Traits related to Conscientiousness (Orderliness; Industriousness) are expected to be negatively correlated with test scores, consistent with the observation that higher levels of Conscientiousness are more prevalent in those of lower ability due to the necessity for these people to develop compensating strategies to effectively meet academic and work goals (Chamorro-Premuzic & Furnham, 2004). Traits related to Agreeableness (Compassion; Cooperation; Sensitivity) are expected not to show any relationship with test performance as no consistent links – either conceptual or empirical – have been observed between these traits and test performance (Chamorro-Premuzic & Furnham, 2004).

In addition to these linear predictions, it is predicted – in two particular cases – that the relationship between Trait scale scores and test performance will be better described by curvilinear relationships. The first of these is the relationship between Stability and test performance, which is expected to adhere to an approximate Yerkes-Dodson curve (e.g. Le Fevre, Kolt & Matheny, 2006). Therefore, it is likely that a quadratic curve will be able to explain more variance than a linear one for this relationship. Similarly, there is a degree of evidence to suggest – albeit in the context of academic performance – that very high levels of optimism can have a detrimental effect on performance (e.g. Furnham, Chamorro-Premuzic & McDougall, 2003). Though, for the most part, optimism is likely to be beneficial to test performance, the over-confidence represented by extreme optimism might lead some participants to expect correct answers, making them less likely to notice the nuance in some

questions. Therefore, this relationship might be better explained by a quadratic or cubic regression equation than by a linear one.

H_8: Socio-economic variables will be able to account for meaningful variance in scale scores for these personality traits.

Based on the findings of Jonassaint et al. (2011) that lower SES is associated with lower scores on all Big Five traits with the exception of Agreeableness, it is expected that socio-economic variables will be able to account for meaningful variance in scale scores for traits that map conceptually on to these personality factors. However, personality has been shown to have both a heritable and an environmental component (Bouchard & McGue, 2003). Therefore, a smaller proportion of the variance in personality will be accountable by social status than for test familiarity. It is still expected that substantial variance in ethnic personality differences will be explainable by familial social status, though, as developmental environment has been shown to have marked effects on key personality traits (e.g. Bouchard & McGue, 2003).

H_9: There will be meaningful differences between ethnic groups on key non-cognitive variables in the study.

H_{10}: Test familiarity differences will be able to account for ethnic group test performance differences attributable to g_f.

Based on environmental models of ethnic group test performance differences (e.g. Dickens & Flynn, 2006), it is expected that minority ethnic groups will score lower than the White majority on all socio-economic variables in the study. The fewer opportunities in society afforded to these minority groups will have – through the mechanism of exposure – led to less opportunity to develop test familiarity. These resultant differences in group mean levels

of test familiarity will be able to explain a substantial proportion of the variance attributable to ethnicity, the most accentuated effect of which is likely to be for tests that have a large g_f component.

H_{11}: Key personality traits identified in the previous analyses will be able to explain a proportion of ethnic group test score differences.

It is possible that those traits identified both as being related to performance and as being related to socio-economic background will be able to explain a degree of ethnic group test performance differences due to ethnic group differences in key socio-economic variables due to these groups' reduced opportunity for education, employment and prosperity in society. Though the literature disagrees on the extent to which personality differences exist between ethnic groups, their existence has been confirmed (e.g. Ones & Anderson, 2002; Goldberg, 1998).

H_{12}: A substantial proportion of ethnic group test performance differences will be able to be accounted for by the combined effect of performance facilitating and debilitating factors.

Given that test familiarity is, in theory, an orthogonal construct to personality, it is expected that little variance will be shared between these predictors of test performance. It is, therefore, expected that the consideration of both of these sources simultaneously will be able to explain ethnic group test score differences more effectively than either of them alone.

4.2 Method

4.2.1 Measures

This section will describe the tools used to measure the various constructs of interest in the study. The study involved the administration of five measures of ability, a personality questionnaire, and a demographic questionnaire. This questionnaire was broken into a number of subsections containing tools to measure relevant constructs within the theoretical model for the study (see Figure 11, section 2.5.3) that were expected to predict ethnic test performance differences.

The first measures to be administered to the participants were the Profiling for Success (PfS) range of ability tests. Published by Team Focus, these measures – the PfS Verbal, PfS Numerical and the PfS Abstract – make up a battery of relatively short timed tests that are designed to be administered on-line as part of organisational processes such as selection (Childs et al., 2013). Though there are a number of versions of each, those chosen for the study (Reasoning Skills Level 1) are normed against groups chosen to represent the UK general population. As such, they tend to be used to select candidates applying for entry-level positions, rather than senior and professional ones. These tests were chosen for a number of reasons. Firstly, since the study aimed to examine ethnic test score differences within the UK population, it made sense to use measures that participants might reasonably expect to complete as part of a UK-based selection process. Secondly, their short time limit (12-15 minutes, depending on the test) meant that they could be administered quickly to participants, reducing the total time of the testing session, thus minimising mental fatigue. Finally, these measures all demonstrate sound psychometric properties. Test development data published by Team Focus demonstrates that the PfS tests are all reliable and valid (Childs et al., 2013), so can be seen as robust assessments of the constructs they were designed to measure (see below).

PfS Verbal

The Profiling for Success Verbal Reasoning Skills (PfS Verbal; Childs et al., 2013) measures a candidate's ability to understand written information and determine what follows logically from this information. It consists of 44 questions based on 11 passages of text (4 questions pertaining to each passage). In each question, candidates must judge the truth of a statement based on what is said in the passage. Candidates are presented with three answer options for each question. Candidates can either judge the statement as 'True' or 'False' using deductive reasoning based on the logic of the passage, or that they 'Can't Tell' whether the statement is true or false as the passage does not contain the necessary premises on which to judge its veracity. The test contains three practice items before it begins. Once the practice items have been completed, candidates have 15 minutes to answer as many of the questions as they can, after which time the test will end, regardless of whether they have finished or not.

Raw test scores are compared against the General Population norm group (N = 2930). The test demonstrates good test-retest reliability (r = .73). The test shows good levels of discriminant validity between itself and the other tests in the range designed to measure separate specific abilities, and good levels of convergent validity with the verbal reasoning component of the GMAT (Childs et al., 2013). The PfS Verbal shows fairly consistent White-Non-white group test score differences of between 0.40 and 0.98 S.D. units, dependent on level. The Verbal Reasoning Skills Level 1 test shows a White-Non-white mean difference of 0.98 S.D. units (Childs et al., 2013).

PfS Numerical

The Profiling for Success Numerical Reasoning Skills (PfS Numerical; Childs et al., 2013) measures a candidate's ability to use numerical information to solve problems. It contains 40 questions, the content of which is split between mathematical operations, extraction of

data from tables, and interpretation of graphs. All test items were designed to present these numerical problems in the context of everyday situations, such as calculating the total value of sales within a retail department, calculating the required mass of an ingredient in a recipe, and so on. There are five answer options for each question. As for the PfS Verbal, candidates have 15 minutes to answer as many questions as they are able, after which, the test immediately ends. The test's instructions say that calculators are not allowed for the test, though the test itself has an inbuilt calculator function that candidates may use for some of the more difficult questions.

Raw test scores are compared against the General Population norm group (N = 1287). The test's test-retest reliability is acceptable (r = .70). The test shows good levels of discriminant validity between itself and the other tests in the range, and good levels of convergent validity with the numerical reasoning component of the GMAT (Childs et al., 2013). When compared to the PfS Verbal, the PfS Numerical shows smaller mean White-Non-white group differences, amounting to between 0.06 and 0.44 S.D. units, dependent on level. The Numerical Reasoning Skills Level 1 test shows a White-Non-white mean difference of 0.32 S.D. units (Childs et al., 2013).

PfS Abstract

The Profiling for Success Abstract Reasoning Skills (PfS Abstract; Childs et al., 2013) measures a candidate's ability to identify patterns in abstract shapes, identify the rules that govern them, and make decisions on the basis of these rules. The test consists of 70 questions, all of which are presented in an identical format (shown in Figure 22 below).

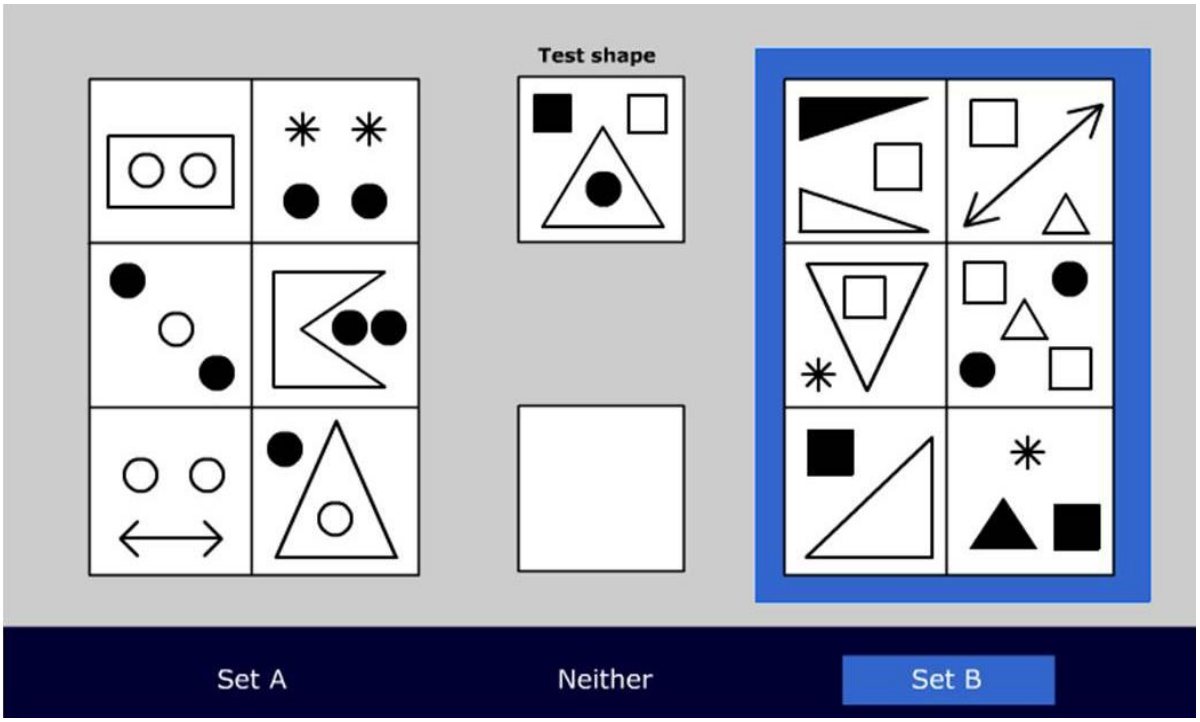

Figure 22. Example item from the PfS Abstract (used with kind permission from Team Focus Ltd.).

For each item, candidates are shown a screen in the format of that shown in Figure 22. Candidates are presented with two sets of six figures in a 2x3 array and a test shape. All of the figures within a set are related according to some implicit rule. First, the candidate must work out the rule that governs the shapes in Set A, and that which governs the shapes in Set B. Once they have done this, they must decide whether the test shape belongs in Set A, Set B, or neither set on the basis of these implicit rules. In the example in Figure 13, all the shapes in Set A contain exactly two circles (irrelevant of whether these circles are coloured black or white). In Set B, all shapes contain at least one square (again, irrelevant of colour). The test shape only contains one circle (meaning that it cannot belong to Set A), but contains two squares. Therefore, the test shape belongs to Set B, so candidates would select this answer option. As with the Raven's tests, the rules that candidates need to apply to each problem are not fixed for the duration of the test. For example, the colour of the shapes is irrelevant in the example shown in Figure 22. However, this is not always the

case. Rather, candidates might have to apply one of a number of rules to each item, or a combination of these rules. Candidates have 12 minutes to complete as many of the 70 questions as they can, after which the test session automatically ends.

Raw test scores are compared against the General Population norm group (N = 455). The PfS Abstract demonstrates slightly less test-retest reliability than the other PfS measures (*r* = .67). The test shows good levels of discriminant validity between itself and the other tests in the range designed to measure separate specific abilities, and between itself and the verbal and numerical sections of the GMAT (Childs et al., 2013). The PfS Abstract again shows smaller mean White-Non-white group differences than the PfS Verbal does, amounting to between 0.02 and 0.55 S.D. units, dependent on level. The Abstract Reasoning Skills Level 1 test shows a White-Non-white mean difference of 0.40 S.D. units (Childs et al., 2013).

Memory and Attention Test (MAT)

As for the PfS range, the Memory and Attention Test (MAT; Team Focus, 2011) is published by Team Focus. It measures six distinct aspects of a candidate's cognition, though its main focus is the measurement of a candidate's ability to memorise increasingly complicated instructions, and apply these quickly and accurately.

The MAT consists of 50 trials, each trial consisting of a test screen to which a candidate must respond. Before each block of five trials, candidates are given a set of instructions that they will be required to apply to each of the subsequent trials. After each block of five trials, a new set of instructions is presented to the candidates, consisting of all the previous instructions, plus an additional rule that they will be required to follow (making each successive block of trials necessarily more complex and cognitively demanding than the preceding ones). An example of the kind of test screen that candidates might be presented with as part of one of these trials is shown in Figure 23 below.

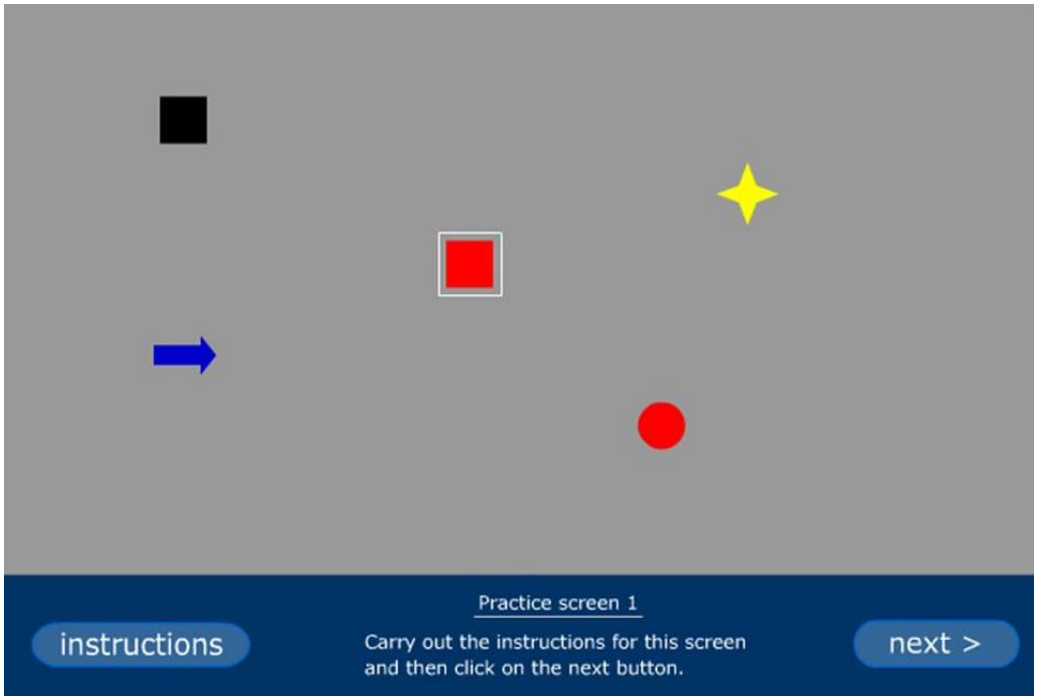

Figure 23. Example item from the Memory and Attention Test.

For the trial shown in Figure 23, candidates might have been given instructions on the previous screen to click on all squares. If this were the case, candidates would need to select the red square and the black square before they clicked 'next', ending the trial.

On the basis of how a candidate performs across all of the 50 trials in the test, scale scores for six dimensions are generated. These six dimensions are accuracy, speed of working, click speed, memory, decisiveness, and decision efficiency. Accuracy is a measure of the number of trials answered correctly (i.e. total score). Speed of working is measured by the total time a candidate takes to complete the test. Click speed (also referred to as 'baseline response') is a measure of how quickly candidates respond to the items in terms of their manual dexterity, reaction time, and ability to use a computer mouse, and is measured at the start of the test in the first few trials. In these trials, the instructions given to candidates are to, for example, "click on all red shapes". As such, these trials are very easy, representing a measure of reaction time as opposed to one of memory or attention. Memory is measured by the number of times a participant needs to recheck the instructions for each trial, and is

recorded by the number of times they click to open the pop-up window containing the trial's instructions. Decisiveness is measured as the number of times a participant changes his or her mind about the correct shapes within the trials, and is measured by the frequency with which shapes are selected and then deselected. Finally, Decision Efficiency is a measure of working memory capacity (Team Focus, 2011), expressed by calculating the mean number of correct trials per minute.

Raw scores for each of the six measures are compared against the Job Applicant Norm Group (N = 675) and converted to percentile scores to allow them to be interpreted, as for the PfS tests. The MAT demonstrates very good internal consistency (α = .84) based on Accuracy scores. Though validity data for the MAT is relatively sparse at present, given that it is relatively new to the market, it shows good levels of convergent validity with the PfS tests based on a sample of 100 participants. Accuracy has been shown to be moderately correlated with scores on the PfS Numerical (r = .30, p < .05) and the PfS Abstract (r = .32, p < .05). Decision Efficiency shows moderate correlations with the PfS Verbal (r = .24, p < .05) and the PfS Abstract (r = .34, p < .05). Speed of Working shows a moderate correlation (r = .24, p < .05) with the PfS Verbal (Team Focus, 2011).

Raven's Standard Progressive Matrices (SPM)

To ensure a degree of consistency across studies in the research, the testing session included the Raven's SPM. This version of the SPM is identical to the test featured in the first archive of Study 1 (see section 3.2.1 for details).

Trait Personality Inventory

Personality was assessed in the study using the Trait Personality Inventory (Trait). Published by Aston Business Assessments (ABA) in 2011, Trait measures personality according to 13 distinct dimensions that are structured around the Big Five model. It was

specifically designed for use in organisational contexts, so – in addition to scales designed to measure important facets of the Big Five – it contains three scales that measure traits important to success in organisations. The inventory contains 127 items, all of which comprise a common item stem (*"In general, I see myself as someone who…"*), followed by a statement about the candidate's personality (e.g. *"feels comfortable around people"*). Candidates respond to each item by rating their agreement with this statement on a 5-point Likert scale.

Table 45 below contains a summary of the 13 Trait scales. For each scale label, there is a short description that captures the breadth of the construct assessed by the scale.

Trait Scale	Construct Descriptor
Sociability	The extent to which someone seeks and enjoys social interaction with others, how likely they are to initiate and develop social contacts and be comfortable doing so.
Leadership	The extent to which people prefer to take a higher profile in groups, to be socially ascendant, to seek recognition and to lead others to achieve.
Compassion	The extent to which people are interested in, and affected by the problems and feelings of others.
Cooperation	The extent to which people prefer to cooperate with others at work, and to help others without expecting or seeking something in return.
Sensitivity	The extent to which people are attuned to emotions and aware of them in themselves and others.
Optimism	The extent to which people feel generally positive and in control of their world, and people's expectations about their own successes.
Achievement	The extent to which people are ambitious, competitive, and achievement-oriented in respect of goals and objectives.
Orderliness	The extent to which people are organised and rule conscious, prefer to work according to plans, and conduct activities in a methodical and orderly manner.
Industriousness	The extent to which people are reliable, hardworking, and committed to finishing tasks and projects they start.
Stability	The extent to which individuals appear relaxed and carefree, versus anxious, worrisome or apprehensive, particularly in response to pressure or challenges.
Calmness	The extent to which people are generally calm, tranquil, and less bothered by irritation, anger or frustration.
Culture	The extent to which people enjoy new experiences and are generally positive about change, and working in new cultures.
Intellect	The extent to which people are intellectual, and interested and open to abstract or theoretical ideas, or complex problem-solving.

Table 45. Summary of the 13 Trait scales (adapted from ABA, 2011).

Ten of the thirteen Trait scales were designed to capture aspects of the Big Five at the facet level, each Big Five trait being measured by two Trait scales. The remaining three scales – Optimism, Achievement and Sensitivity – were included to capture personality preferences

that related to positive psychology (such as self-efficacy), drive for personal and organisational results, and emotional intelligence respectively.

Based on a candidate's responses to each of the 127 items in the inventory, raw scores are calculated for each of the 13 Trait scales. These raw scale scores are then compared to a norm group that represents the UK working population (N = 1273). Sten scores are generated according to the position of a candidate's score within the distribution of scores for this sample. Inferences are then made about their personality on the basis of sten score bandings ('low', 'average', or 'high').

All of the Trait scales have demonstrated good levels of test-retest reliability (r = .72 – .89) and internal consistency (α = .76 – .85). The Trait scales show good structural validity, loading on to five factors that map closely to the Big Five. The scales all show good convergent and divergent validity between themselves and with the scales of the BFI (ABA, 2011).

Participant and Familial Social Status

Socio-economic status was assessed using measures adapted from Hollingshead's (1975) Four Factor Index of Social Status. It measures social status (a widely-accepted proxy for socio-economic status) as an aggregate measure of level of education, marital status, gender and occupational prestige, the latter being defined using a hierarchical occupational taxonomy. This allows social status scores to be calculated for both individuals and households that are not confounded by differences in the spectrum of modern living situations, such as single-parent families, two-parent families in which only one parent is employed, and those in which both are employed (Hollingshead, 1975).

Though Hollingshead's measure was never published, it has been phenomenally successful as a measure of social status/SES, particularly in the Sociology literature (Adams & Weakliem, 2011). According to Google Scholar, to date (August 2015), it has been cited in

published material no fewer than 8103 times, around 500 of these in the last 12 months alone. As such, it has been – and continues to be – hugely influential on the study and measurement of socio-economic variables in a variety of disciplines.

A criticism that has been levelled at Hollingshead's original measure, though, is that it can be somewhat unwieldy to administer in practice. To address this, more practical, simplified iterations of it have been proposed that retain Hollingshead's underlying theoretical structure, but present it in a way that is more user-friendly for both those administering it and those completing it. One such measure is the Barratt Simplified Measure of Social Status (BSMSS; Barratt, 2006, cited by Adams & Weakliem, 2011). Barratt's measure is presented as a simple two-page self-report questionnaire, comprising a scale to measure educational attainment and one to measure occupational prestige. The practical advantage of the BSMSS over Hollingshead's original tool is in how the taxonomy of occupations is presented. Rather than giving participants an exhaustive list of occupations included in each of Hollingshead's nine ordinal categories, the BSMSS presents each category as a list of example occupations that capture the breadth of the category. This allows the scale to be presented in a much more efficient way than in its original form.

In addition to this presentational modification, Barratt makes two important changes to Hollingshead's index. Firstly, the taxonomy of scales has been slightly adjusted based on updated data from the work of Davis et al. (1991), meaning that its assessment of an occupation's relative prestige is based on somewhat more recent data than the original tool. Secondly, it recognises the influence that a person's social background has on their social status by capturing the occupational prestige and educational level not only of the participant completing the measure, but also of their mother and father while they grew up.

The present study used a measure of Social Status that was loosely based on the BSMSS. Hollingshead's index, and, by extension, tools which use it as a foundation such as the BSMSS, have been criticised for their focus on occupational *prestige* as opposed to income

(e.g. Hauser & Warren, 1997). It is recognised that this measure of social status does not fully capture SES as it is normally defined (e.g. APA Task Force on Socioeconomic Status, 2007), but controlling for chronological differences (due to inflation) and national differences (due to the relative wealth of country of residence) in a participant's familial income and wealth would have been extremely difficult (if not impossible). Therefore, it was judged that this measure would capture the socio-economic background of participants to a sufficiently accurate degree for the purposes of the present study.

For the present study, two minor changes were made to the BSMSS and how it is used. Firstly, the taxonomy was changed slightly from the original US English naming of occupations to be more in line with UK occupational naming conventions (though the hierarchy of the occupations itself remained unchanged). Secondly, participants were asked to rate their own present educational level, and that of their mother and father at the time during which they attended the majority of school. Rather than being used, as in the BSMSS, to form a composite measure of an individual's social status, these measures were used to capture the separate constructs of one's current social status and their socio-economic background.

Using data obtained by these scales, two ordinal variables were generated for each participant, one reflecting the participant's present social status and the other reflecting the social status of their family during their formative years. Familial access to education was scored according to participants' responses on a 7-point scale of educational level for their mother and father, the mean of these two scales being taken as each participant's familial access to education score (giving a range of scores from 1–7). Familial access to employment was scored according to participants' responses on a 9-point scale of occupational prestige for their mother and father, the mean of these two scales being taken as each participant's familial access to education score (giving a range of scores from 1–9). Familial social status was calculated using only the scores for each participant's mother and father. This gave a range of scores from 6–66, the scores being weighted to reflect

Hollingshead's (1975) original weighting of 5:3 in favour of occupational prestige over educational level, recognising the larger contribution of occupational prestige to social status. The same procedure was used to calculate Participant Social Status, using the educational level and occupational prestige of participants in place of those of their parents.

Other Demographic Variables

In addition to social status, a number of key demographic variables were captured by an on-line questionnaire hosted by Survey Monkey. The questionnaire recorded each participant's age in years, gender, ethnicity, employment status, job title (if applicable), approximate number of hours worked per week (if applicable), country of birth, country of residence during development, and country in which the majority of school was attended. Based on guidelines for ethical data collection laid out by both the BPS (2010) and the ESRC (2012), participants were given the right to refuse to answer any question within the questionnaire. Responses to the questionnaire were kept anonymous through use of unique participant ID numbers (see section 4.2.3), which participants were asked to enter on the first screen of the questionnaire.

4.2.1.1 Designing the Test Familiarity Scale

As test familiarity is a key variable in the theoretical model described in the previous chapter, the questionnaire additionally contained a five-item scale designed to measure each participant's familiarity with ability testing. Currently, within the literature, no scale exists to assess test familiarity. Therefore, in order to assess test familiarity in a robust way, a scale needed to be developed to capture this construct. The aim of this process was to develop a concise, psychometrically robust scale of between 3 and 5 items that could be used to obtain a quick, reliable measure of a participant's test familiarity.

The scale development process began by defining the construct that the scale should capture. The conceptualisation of test familiarity used was based on a revised version of Reeve, Heggestad and Lievens' (2009) definition, that characterised it as a broad construct, not specific to a single test, but generalisable across tools, and that was distinct from practice effects in that it is somewhat persistent over time. The definition of test familiarity used for the process of scale development, therefore, was as follows:

> *"Test familiarity encompasses all construct-irrelevant test knowledge and skills that a test taker has accumulated over time through exposure in both organisational and wider, non-organisational contexts, that are not associated with any general or specific ability, yet facilitate performance on tests designed to measure them."*

Scale items were then generated using a deductive approach to scale design (e.g. Burisch, 1984; Hinkin, 1995). By utilising this approach, fewer items needed to be generated than had an inductive approach to scale development been taken (Burisch, 1984), since the scale's construct was clearly defined based on a thorough review of the test familiarity literature (see section 2.5.2). On the basis of the construct's definition, five items were generated to make up the scale, using a mixture of positively- and negatively-worded items to avoid response pattern biases (Idaszak & Drasgow, 1987). Each item required participants to rate their strength of agreement with a statement about their degree of test familiarity on a 5-point Likert scale. The items generated are shown below:

1. *"Before today, I was familiar with ability tests"*
2. *"Generally, I understand the questions asked in these tests"*
3. *"I hadn't taken many ability tests before today"*
4. *"I have experience of taking ability tests when applying for jobs"*
5. *"I'm not very good at working out how to answer the questions asked in these tests"*

These items all aimed to tap into the construct, as defined above, to some extent. As well as containing items that directly assessed a participant's test familiarity (Items 1, 3, and 5), both in general and in the context of selection, it was reasoned that the component of test familiarity that relates to facilitation of performance should also be represented. To that end, Items 2 and 5 relate to accumulated test knowledge and skills, in that they assess the degree to which participants are able to understand the content of test items and what is required to solve them, irrelevant of their level of ability.

Once the item set had been generated, it was validated in a pilot study of 59 participants (a subset of the study's main sample). The item responses collected for these participants were recoded so that all responses were positively keyed. The scale was then developed using an approach in which decisions on whether to retain or reject items in a pool for the final scale are made on the basis of a number of criteria. Firstly, an item may be flagged for deletion if its removal would meaningfully increase the scale's overall value of Cronbach's α. Secondly, when all items in the initial scale are entered into an EFA with oblique rotation in which a two-factor solution is forced, if an item designed to be part of a unifactorial scale (as the present scale was intended to be) clearly marks out a second, nuisance factor, that item may be flagged for deletion. Thirdly, when a single-factor solution is forced in the same analysis, if any item shows either low loading on that single factor or low communality at extraction, that item may be flagged for deletion. Based on this process, the number of 'flags' for each item are examined, and a decision made whether to accept or reject its inclusion in the final scale by assessing the impact that either course of action would have on how the scale functions.

The scale showed acceptable internal consistency (α = .76). This scale α could only be improved by the removal of one item, Item 5, but its removal only increased the value of α by .02. To test the structural validity of the new scale, EFA was conducted in Mplus using diagonally weighted least squares (WLSMV) estimation and PROMAX rotation. These options were chosen to accommodate the ordinal nature of the data, allowing Mplus to

generate polychoric correlations. More specifically, WLSMV has been demonstrated to outperform other EFA estimation methods when used with ordinal data at smaller sample sizes (Bandalos, 2014; DiStefano & Morgan, 2014). The eigenvalues and resultant scree plot (see Figure 24) generated by this analysis both pointed to a unifactorial solution. This implies that a single factor underlies the scale. This was confirmed through PA-MRFA in FACTOR, the first extracted factor explaining 61.0% of the variance with an eigenvalue of 3.03.

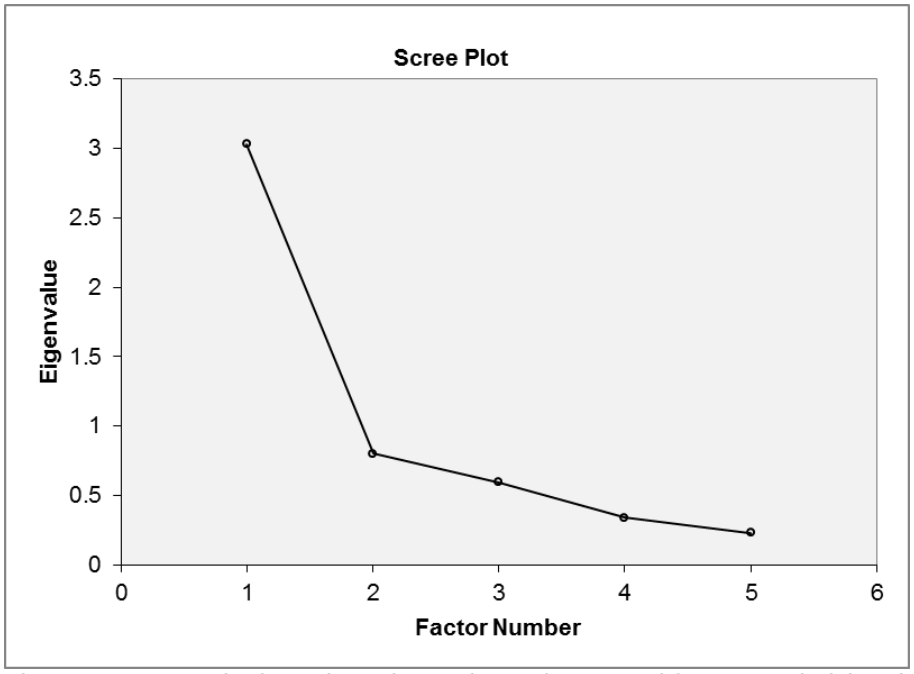

Figure 24. Scree plot based on eigenvalues of extracted factors underlying the test familiarity scale (pilot study).

The results of the EFA conducted on the items within the scale showed that all items loaded meaningfully on to a single underlying factor, showing loadings greater than .3 (Hair et al., 1998). All communalities for the items were at acceptable levels, with the exception of Item 5 (showing a communality of .17). This and the reliability analysis suggested that the scale might benefit from the deletion of Item 5. However, since its removal would represent only a marginal incremental improvement to the scale, it was judged to be prudent to reassess this once the total sample had been collected.

The full scale was administered to participants at the end of the demographic questionnaire. On the basis of the response data collected for the full sample (N = 126; 7 missing), the psychometric properties of the scale were again assessed. The scale showed an acceptable value of Cronbach's α (.70). Contrary to the findings of the pilot study, this value did not improve with the deletion of any of the items. EFA with WLSMV estimation once again suggested a unifactorial structure underlying the items on the basis of both its eigenvalues and scree plot (see Figure 25). This was confirmed through PA-MRFA, the first extracted factor explaining 63.5% of the variance with an eigenvalue of 2.54.

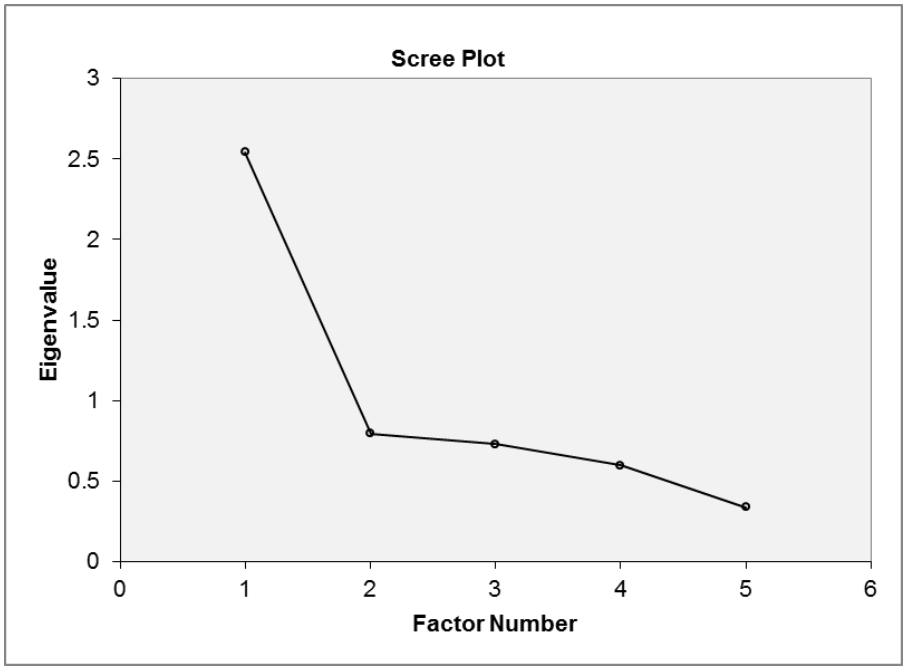

Figure 25. Scree plot based on eigenvalues of extracted factors underlying the test familiarity scale (full sample).

All items showed loadings above .4 on this single factor. However, Item 5 again showed poor communality with the other items (.19), and showed the smallest loading of the five items (.44) onto the single factor. On the basis of these results, it was judged to be prudent to revisit the wording of this item to ensure that it effectively captured the construct as defined.

Re-examining Item 5, it is entirely possible that the item's wording – *I'm not very good at working out how to answer the questions asked in these tests* – is causing it to tap into a construct that is highly correlated with test familiarity, but is not test familiarity in and of itself. It is entirely possible for a respondent to be extremely familiar with ability testing, but self-aware of their own limitations enough to know that they are no good at them. As such, this item more likely taps into Chamorro-Premuzic and Furnham's (2004) subjectively assessed intelligence (SAI) construct than it does test familiarity. By the same rationale, it could be argued that Item 2 – *Generally, I understand the questions asked in these tests* – could also tap more into SAI than it does test familiarity. However, the distinction lies in the difference in focus between these two items: Item 5 focuses on a respondent's confidence in their ability to *answer* test questions, which depends on ability (or, rather, their *perception* of their own ability). Conversely, Item 2 examines a participant's *comprehension* of what a question is asking them, irrelevant of whether or not they possess the level of ability necessary to be able to answer it correctly.

Based on this reasoning, a rational decision was made to remove Item 5 from the final scale, and to keep Item 2. EFA was rerun on the remaining four items in the scale. All four items loaded strongly onto a single factor, and displayed acceptable communality (>.30) with the other items. Cronbach's α for the four-item scale was calculated as .69. Though this is slightly below the rule of thumb for acceptable α values of .70, this was judged to be acceptable, given the small number of items in the scale. Removal of further items did not improve the scale's α value.

The final stage in the scale development process was to assess the scale's concurrent validity. To do this, the association between test familiarity and test performance was examined. Participants' mean item scores were calculated, and this new variable was used to represent their test familiarity (as opposed to as a sum of their item scores as any missing data would have skewed scale score had this approach been used). The variable was then correlated with mean raw scores for all the measures of ability used in the study. The new

scale was significantly correlated with all measures of ability (with the exception of the MAT's Speed of Working scale), showing correlation coefficients between .19 (MAT Accuracy; $p < .05$) and .28 (PfS Numerical; $p < .01$). This suggests that the scale likely measures what it was designed to measure, higher test familiarity being associated with better performance on ability tests across the board.

4.2.2 Sample

4.2.2.1 Sampling Strategy

The study adopted a combination of approaches to the recruitment of participants. In the initial phase of data collection, participants were primarily recruited through the author's personal and professional networks using convenience sampling. In the second phase, snowball sampling was employed by recruiting 11 participant researchers (all of whom held the BPS' Occupational Test User: Ability and Personality certificates) and training them to recruit ten new participants and administer the measures to them in proctored test conditions. The population from which the participant researchers were instructed to draw the sample was UK residents of working age (though, to avoid range restriction, participants were not required to be employed). This phase also made use of stratified sampling to a degree, as these participant researchers were instructed to recruit ten participants who, as a group, represented an ethnically and socially heterogeneous sample of this broad population. Towards the end of this phase, the stratification of the sampling became more rigorous, later participant researchers being instructed to recruit participants from narrow ethnic groups, in order to try to ensure the representativeness of the total sample. In the third phase of the data collection, the author contacted a number of educational establishments and other organisations directly in an attempt to recruit participants *en masse*. Of the organisations that were contacted in this phase, those able to provide the greatest number of participants were a further education institution in the West Midlands and

a higher education institution in the East of England. In spite of this tripartite approach to data collection, participant response rates were extremely poor throughout the data collection phase of the study.

4.2.2.2 Conceptualising Ethnicity

For the purposes of the present study, ethnicity was defined according to the acculturation model defined by Phinney (1990). This working definition conceptualises ethnicity as *"not only* [concerned with] *the strength of ethnic identity but also* [with] *the relationship to the majority culture"* (Phinney, 1990, p.510). As such, participants were asked to report not only their ethnicity according to one of a number of ethnic categories that mirrored those of the ONS' (2013) standard question, but were also asked – if applicable – to rate the strength of identification with both UK culture and the culture of any country in which they might have lived during their formative years.

Within the demographic questionnaire, participants were asked to report the country of their birth, the country in which they spent the majority of their developing years, and the country in which they attended the majority of school. If any of these were outside the UK, participants were asked to rate the strength of their identification with both the culture of the country they reported in these previous questions and UK culture. This allowed for richer interpretations of the ethnic categories reported in the participants' responses to the ONS' standard question.

According to their cultural identification ratings, participants were classified as belonging to one of four categories, dependent on their strength of identification with UK culture and the culture of their country of birth, development and/or schooling. This classified participants as being either acculturated (strong UK, strong ethnic), assimilated (strong UK, weak ethnic), dissociated (weak UK, strong ethnic) or marginal (weak UK, weak ethnic). For the purposes of the study, British participants were coded as being assimilated. Participant ethnicity as

measured by the standard question was then recoded on the basis of this cultural identification (see section 4.2.2.4).

In addition to this method of classifying ethnicity, the study sought to examine ethnic group differences at the levels of abstraction normally used in ethnic bias studies of this kind. To this end, ethnicity was further recoded into two separate variables. In the first, participants were recoded as being either 'White' or 'Non-White'. For the second, ethnicity was again recoded into the broad ethnic classifications that tend to be referred to in the literature (i.e. 'White', 'Black', 'Asian', etc.). However, this second variable was manipulated in a number of ways in an attempt to overcome some of the problems of ethnic classification highlighted by Study 1. On the basis of their strength of ethnic identification, any participant who had been identified as ethnically non-British was excluded from this variable. This was done primarily to control for the possible confounding effect of the inclusion of non-British White test takers in the White group, given that these test takers tend to perform substantially better than their White British counterparts. Similarly, Chinese participants and those who had reported their ethnicity as 'Any other Asian ethnic group' were excluded from the Asian group. The resultant ethnic group became a broadly homogeneous representation of Asian ethnic subgroups whose origin was within the Indian Subcontinent, as these were the most populated of the Asian subgroups.

4.2.2.3 Conceptualising Employment Status

A participant was deemed to be unemployed if they satisfy the criteria adopted by the International Labour Organisation (ILO, 2004). According to this definition, a person may be classified as unemployed if they are a) above a certain age, b) without work, c) currently available for work, and d) seeking work.

Employment status was, therefore, represented as a polytomous nominal variable. If a participant satisfied all of the above criteria, their employment status was coded as 0

(unemployed). If a participant did not satisfy all of these 4 criteria but was not employed, their employment status was coded as 1 (without employment). Employed participants were coded as 2 (employed).

4.2.2.4 Participants

Data for the study were collected from a total sample of 133 participants. Ninety-nine participants (74.4%) described their gender as 'female', and the remaining 34 (25.6%) described their gender as 'male'. Participants' mean age was 25.32 years (S.D. = 10.74 years; Range = 16–66 years). The highest level of qualification obtained by participants was reported to be 'GCSE or equivalent' by 20 participants, (15.0%), 'some further education' by 3 (2.3%), 'A-level or equivalent' by 39 (29.3%); 'some higher education' by 14 (10.5%), 'first (undergraduate) degree or equivalent' by 15 (11.3%), and 'postgraduate degree or equivalent' by 24 (18.0%), with 18 participants choosing to withhold this information.

Participants were drawn from a wide range of occupational backgrounds. According to the guidelines of the ILO (2011), 3 participants (2.3%) were classified as 'unemployed'. Forty participants (30.1%) were classed as Further Education students. Twenty-eight participants (21.1%) were classed as full-time Higher Education students. Only one participant (0.8%) did not report their occupational status. Of the remaining 61 employed participants (46.0%), 26 (19.5%) worked in business occupations, 9 (6.8%) in medical occupations, 8 (6.0%) in the service industry, 7 (5.3%) in academia, 5 (3.8%) in retail, 4 (3.0%) in social care occupations, and 2 (1.5%) in the entertainment industry.

As was to be expected from a UK-based study, the vast majority of the sample were born (109; 82.0%), schooled (117; 88.0%), and/or spent their developing years (115; 86.5%) in the UK. The remaining participants were distributed in the following ways: Two participants each (1.5%) reported their country of birth as Germany, Romania or the USA, and a single participant (0.8%) reported their country of birth as being India, Pakistan, Bangladesh,

Malaysia, Kuwait, New Zealand, Kenya, Nigeria, Poland, Latvia, or the Seychelles. Two participants (1.5%) reported their country of development as Romania, and a single participant (0.8%) reported theirs as being India, Malaysia, Germany, USA, Kuwait, New Zealand, Poland, Latvia or the Seychelles. Two participants (1.5%) reported the country in which they attended the majority of school as Romania, and a single participant (0.8%) reported theirs as being Malaysia, Germany, Kuwait, New Zealand, Poland or Latvia. The cultural identification ratings of these non-UK participants were then examined to determine the bearing that this had on the interpretation of their reported ethnic group. Nine participants were found to identify much more strongly with UK culture than with that of their countries of birth, development and/or schooling, so were classed as being ethnically British. The remaining participants identified much more strongly with the culture of the country of their birth, development, and/or schooling, so were classified as ethnically non-British.

On the basis of this, 63 participants were classified as 'White – British' (47.4%), 8 as 'Any other White background' (6.0%), 6 as 'Black British – African' (4.5%), 6 as 'Black British – Caribbean' (4.5%), 18 as 'Asian British – Indian' (13.5%), 17 as 'Asian British – Pakistani' (12.8%), 2 as 'Asian British – Chinese' (1.5%), 1 as 'Asian British – Bangladeshi' (0.8%), 2 as 'Asian British – Any other Asian background' (1.5%), 3 as 'Black British – Any other Black background' (2.3%), 1 as 'Middle Eastern' (0.8%), 2 as 'Mixed – White and Asian' (1.5%), 1 as 'Mixed – White and Black Caribbean' (0.8%), 1 as 'Mixed – White and Guyanese (0.8%), and 2 (1.5%) as 'unreported ethnicity'. Finally, these revised ethnic classifications were conflated into three broad ethnic groups (following the process described in section 3.2.2.2). Within this variable, there were 63 White British participants (55.3%), 15 Black British participants (11.3%) and 36 Asian British participants (27.1%).

4.2.3 Procedure

Once they had been recruited for the study, participants were sent three e-mails. At the same time, their e-mail addresses were entered into Team Focus' assessment platform and a unique participant ID number was generated for each of them. The first e-mail thanked participants for their help and instructed them to go to ABA's website to complete the Trait Personality Inventory prior to the testing session. Additionally, they were instructed not to open and follow the links in the other two e-mails until they were in the testing session itself and had been told to do so by the session's administrator.

At the start of the testing session, participants were welcomed and logged into a computer. First, it was ensured that participants had all received the three e-mails and had completed Trait. If they had not completed Trait, they were told to do so as soon after the testing session as they were able. If they had not received the 2nd and 3rd e-mails (those containing the links to the ability tests and demographic questionnaire), these were immediately resent to them. Participants were each given the unique participant ID number that had been generated for them. They were also given a pen and two sheets of paper on which to make notes and rough calculations.

A letter of consent was then distributed to participants, which they we instructed to read, and sign if they were happy to participate in the experiment. This letter of consent talked in general terms about the nature of the study (e.g. *"The study is designed to help us to understand how different groups of people might approach answering ability test questions in different ways."*) but its wording was deliberately vague, particularly with reference to the study's focus on ethnic group test performance differences. The reason for this was to avoid the possibility of stereotype threat confounding the results (see section 2.5.1). Had ethnicity been made salient in the testing session, either explicitly (by making participants aware of its existence), or implicitly (by asking them to report their ethnicity) prior to testing, the study's results could have been confounded by stereotype threat. Therefore, to avoid this, it was

ensured that ethnicity was not made salient in the study by revealing its true intentions prior to administration of the tests.

If any participants were unwilling to participate, they were thanked and told to leave. The signed letters of consent were then collected for the remaining participants. The testing session then proceeded according to the best practice guidelines for ability test administration laid out as part of the BPS' Occupational Test User: Ability qualification (BPS, 2011). This allowed the testing environment to be standardised as much as possible to avoid error variance in test scores due to differences in the method of administration across testing sessions.

Participants then completed the tests in the order of the PfS Verbal, the PfS Numerical, the PfS Abstract, the MAT, and, finally, the Raven's SPM. This order of presentation was kept consistent to avoid the influence of order effects due to fatigue on later tests. Participants were given a short comfort break between administration of the PfS Abstract and the MAT. During this comfort break, any rough paper used in the preceding tests was collected so that participants were not tempted to make use of it as an aid to memory during the MAT.

Once they had completed all the ability tests, a letter of debriefing was distributed to each participant. At that point in the testing session, there was no way that stereotype threat could have influenced their ability test scores, so this letter described the nature and purpose of the study in explicit detail. Participants were reminded of their right to withdraw from the research at any time, particularly if they felt uncomfortable now that they knew the true purpose of the study.

Once they had completed and signed the letter of debriefing, these were collected and participants were instructed to follow the final link in the e-mail, which directed them to a short demographic questionnaire. The researcher was on hand while they completed this to answer any questions they had about specific questions within the questionnaire, as well as to answer more general questions about the study.

Once they had completed this final part of the testing session, participants were thanked and told they were free to leave. From start to finish, the whole testing session took between 2 hours, and 2 hours and 30 minutes.

4.3 Results

The separate data files from Talent Lens', Team Focus', and Survey Monkey's online assessment platforms were downloaded in Excel '97–2003 (.xls) format. These files were then merged into a single master file. These data were recoded and then imported into SPSS version 20 for analysis.

There were some instances of missing data in the dataset. This is largely due to the approach taken to ensure that data were collected ethically from participants (see section 4.2.1). When responding to the questionnaire, participants were given the right to refuse to answer any question without providing a reason for doing so. While most participants provided answers to all questions within the questionnaire, there were some instances in which questions were skipped for whatever reason. This created the dilemma of how best to handle this missing data. There are a number of opposing viewpoints as to the most appropriate method by which to address missing data in a dataset, be it listwise deletion, single imputation (both of which are discussed at length in Little & Rubin, 1987), multiple imputation (Rubin, 1987), or one of a growing number of more complex statistical approaches (e.g. Efron, 1994). Given that the missing data were randomly distributed within the dataset, listwise deletion of data was judged to be the technique least likely to introduce bias into the present study's dataset. Though listwise deletion necessarily reduces statistical power as it reduces the sample size used for analysis, this was not judged to be overly problematic, given the relatively small incidence of missing values.

4.3.1 Ethnic Mean Test Score Differences

To address Study 2's hypotheses, the first step was to determine whether ethnic group differences in test performance existed for these measures, as had been demonstrated for the Raven's tests in Study 1. Mean raw scores were calculated for the four measures of ability (PfS Verbal, Numerical, and Abstract, and Raven's SPM), and for the three subscales within the MAT that had been shown to be most strongly related to ability (i.e. Accuracy, Decision Efficiency, and Speed of Working). It is worth pointing out at this stage that scores on the Speed of Working subscale are calculated as the sum of the time taken to complete all trials in the test. As such, a lower score indicates a quicker speed of working.

Mean group raw scores on each of the ability measures were calculated for the two conceptualisations of ethnicity used in the study, along with their standard deviations, and Cohen's *d* in comparison to the White majority group. These results are shown in Tables 46 and 47 below.

Ethnic Group		White	Non-White
PfS Verbal			
	N	72	56
	Mean Raw Score	28.21	22.64
	(S.D.)	(6.49)	(5.96)
	Cohen's *d*	-	0.89
PfS Numerical			
	N	72	55
	Mean Raw Score	23.83	22.45
	(S.D.)	(4.03)	(2.92)
	Cohen's *d*	-	0.39
PfS Abstract			
	N	72	56
	Mean Raw Score	41.33	37.84
	(S.D.)	(12.34)	(11.69)
	Cohen's *d*	-	0.29
Raven's SPM			
	N	71	54
	Mean Raw Score	18.82	15.31
	(S.D.)	(5.55)	(5.91)
	Cohen's *d*	-	0.61
MAT			
	N	71	56
	Accuracy	28.92	24.21
	Decision Efficiency	3.23	2.88
	Speed of Working	547.21	540.29

Table 46. Mean raw scores and Cohen's *d* by ethnicity classified as White/Non-white.

Ethnic Group		White	Black	Asian
PfS Verbal				
	N	63	15	36
	Mean Raw Score	28.37	20.67	23.1
	(S.D.)	(6.52)	(6.32)	(5.81)
	Cohen's *d*	-	1.20	0.85
PfS Numerical				
	N	63	15	36
	Mean Raw Score	23.56	20.93	22.81
	(S.D.)	(3.93)	(4.27)	(2.30)
	Cohen's *d*	-	0.64	0.23
PfS Abstract				
	N	63	15	36
	Mean Raw Score	40.59	32.0	40.06
	(S.D.)	(12.24)	(8.18)	(11.58)
	Cohen's *d*	-	0.83	0.04
Raven's SPM				
	N	62	15	35
	Mean Raw Score	18.31	16.87	14.89
	(S.D.)	(5.58)	(4.98)	(6.24)
	Cohen's *d*	-	0.27	0.58
MAT				
	N	62	15	36
	Accuracy	28.63	22.67	24.86
	Decision Efficiency	3.21	2.65	2.81
	Speed of Working	546.27	555.80	534.72

Table 47. Mean raw scores and Cohen's d by broad ethnic group.

Examining these results in their entirety, it appears that there are persistent ethnic group differences in performance across different tests designed to measure different facets of ability. There was a degree of variability to the findings in terms of the magnitude of ethnic differences, values of Cohen's *d* between specific ethnic groups and the majority group ranging from around 0.3 S.D. units to a full S.D. unit. These differences seemed most pronounced for the PfS Verbal, though it appears that in almost all cases, the White majority group substantially outperformed those participants from minority ethnic groups. Furthermore, the pattern of differences observed between White and Non-white groups for the PfS tests closely matches those observed by Childs et al. (2013) during the test battery's validation, suggesting that these results are representative of the typical level of performance for these groups.

These findings, then, would appear to lend support to H_1. However, these differences first need to be contextualised in terms of the variance in total test score that ethnicity can explain.

4.3.2 The Variance in Test Scores Explainable through Ethnicity

A series of multiple regression models were run to determine the extent to which ethnicity could explain test performance across different measures of ability. In each model, the multinomial ethnic classification variables were dummy coded into dichotomous variables to represent each separate ethnic category. Total raw score for each measure was then regressed on to these dummy variables. The results of these analyses are shown in the following tables.

PfS Verbal

	B	S.E. B	*β*
Constant	22.64	0.84	
Ethnicity	5.57	1.11	0.41*

Table 48. Linear regression of ethnicity reclassified as White-Non-white predicting total PfS Verbal raw score. Note: R^2 = .17; *p < .001

	B	S.E. B	*β*
Constant	28.37	0.79	
Black	-7.70	1.80	-0.38*
Asian	-5.31	1.31	-0.36*

Table 49. Linear regression of broad ethnic group predicting total PfS Verbal raw score. Note: R^2 = .20; *p < .001

PfS Numerical

	B	S.E. B	*β*
Constant	22.455	0.48	
Ethnicity	1.379	0.64	0.19*

Table 50. Linear regression of ethnicity reclassified as White-Non-white predicting total PfS Numerical raw score. Note: R^2 = .04; *p < .05

	B	S.E. B	*B*
Constant	23.56	0.45	
Black	-2.62	1.02	-0.25*
Asian	-.75	0.74	-0.10

Table 51. Linear regression of broad ethnic group predicting total PfS Numerical raw score. Note: R^2 = .06; *p < .05

PfS Abstract

	B	S.E. B	*β*
Constant	37.84	1.61	
Ethnicity	3.49	2.15	0.14

Table 52. Linear regression of ethnicity reclassified as White-Non-white predicting total PfS Abstract raw score. Note: R^2 = .02

	B	S.E. B	*B*
Constant	40.59	1.46	
Black	-8.59	3.33	-0.25*
Asian	-0.53	2.42	-0.02

Table 53. Linear regression of broad ethnic group predicting total PfS Abstract raw score. Note: R^2 = .06; *p < .05

Raven's SPM

	B	S.E. B	*β*
Constant	15.32	0.78	
Ethnicity	3.50	1.03	0.29*

Table 54. Linear regression of ethnicity reclassified as White-Non-white predicting total Raven's SPM raw score. Note: R^2 = .09; *p < .001

	B	S.E. B	*B*
Constant	18.31	0.73	
Black	-1.44	1.65	-0.08
Asian	-3.42	1.21	-0.27*

Table 55. Linear regression of broad ethnic group predicting total Raven's SPM raw score. Note: R^2 = .07; *p < .01

MAT

Accuracy

	B	S.E. B	*β*
Constant	24.21	0.89	
Ethnicity	4.70	1.20	0.33*

Table 56. Linear regression of ethnicity reclassified as White-Non-white predicting MAT Accuracy. Note: R^2 = .11; *p < .001

	B	S.E. B	*B*
Constant	28.63	0.83	
Black	-5.96	1.88	-0.30*
Asian	-3.77	1.37	-0.26*

Table 57. Linear regression of broad ethnic group predicting MAT Accuracy. Note: R^2 = .11; *p < .01

Decision Efficiency

	B	S.E. B	*β*
Constant	2.88	0.14	
Ethnicity	0.36	0.19	0.17

Table 58. Linear regression of ethnicity reclassified as White-Non-white predicting MAT Decision Efficiency. Note: R^2 = .03

	B	S.E. B	*B*
Constant	3.21	0.11	
Black	-0.57	0.25	-0.22*
Asian	-0.40	0.18	-0.21*

Table 59. Linear regression of broad ethnic group predicting MAT Decision Efficiency. Note: R^2 = .07; *p < .05

Speed of Working

	B	S.E. B	*β*
Constant	540.29	18.15	
Ethnicity	6.93	24.27	0.03

Table 60. Linear regression of ethnicity reclassified as White-Non-white predicting MAT Speed of Working. Note: R^2 = .00

	B	S.E. B	*B*
Constant	546.27	16.42	
Black	9.53	37.21	0.03
Asian	-11.55	27.10	-0.04

Table 61. Linear regression of broad ethnic group predicting MAT Speed of Working. Note: R^2 = .00

These analyses revealed, overall, that ethnicity can explain substantial variance in test scores. This supports H_1. Some further important findings can be extracted from these results, however. Firstly, the amount of variance in test scores explainable by ethnicity depends heavily on how ethnicity is conceptualised. In most cases, the more complex conceptualisation of ethnicity into broad ethnic groups was able to explain substantially more variance in total test scores than when ethnicity was reduced to a White-Non-white dichotomy. Secondly, of the MAT subscales, ethnicity appeared less able to explain variance in Speed of Working than in the other two scales (though it could explain substantial variance when regressed on to the reduced ethnic classification variables). Thirdly, the amount of variance in test scores explainable by ethnicity is heavily dependent on the test that is being examined. Across all conceptualisations of ethnicity, the proportion of the variance in PfS Verbal scores that could be explained was far larger than that for the other measures, amounting to between 17% and 20% of the variance in test scores. Variance in PfS Numerical, PfS Abstract and Raven's SPM test scores explainable by ethnicity was variable, though was markedly smaller than for the PfS Verbal, ranging from around 2% of the variance to around 9%, depending on the conceptualisation of ethnicity used. Broadly speaking, this pattern can be seen as providing some support for H_2, based on what one might expect from the literature, given that the Raven's measures (Ashton & Lee, 2006) and verbal reasoning tests in general (Beauducel, Brocke & Liepmann, 2001)

have been observed to be highly *g*-loaded. However, to establish whether these patterns constitute firm support for the Spearman-Jensen Effect, this warrants further investigation through consideration of the degree to which each test loads on to *g* in the present study (see section 4.3.3.3 below).

Having established the existence of robust ethnic differences in the sample, the next objective of the study was to attempt to explain them by examining the extent to which these differences could be accounted for by the non-cognitive variables in the dataset, both in terms of factors that might directly affect test performance, and in terms of their root causes. Before this was possible, however, it was first necessary to explore the nature of these factors and how they might influence performance within the sample as a whole.

4.3.3 The Nature of Performance Facilitating and Debilitating Factors

To examine how the facilitating and debilitating factors identified in the study's theoretical model might impact test performance, a number of analyses were conducted. The following sections detail these analyses, beginning with an examination of the socio-environmental factors measured in the study as potential root causes of test performance differences.

4.3.3.1 Socio-economic Variables and their Influence on Test Performance

To establish the degree to which test performance differences could be accounted for by differences in socio-economic background, a matrix of Pearson correlations was generated between raw scores on all cognitive measures and the socio-environmental variables measured in the study. Though the focus of the theoretical model was primarily on the effect of social status on test performance, the facet-level variables of which it consisted were included for the purposes of more thorough exploration of the nature of these antecedent relationships. This matrix of correlations is shown in Table 62 below.

	PfS Verbal	PfS Numerical	PfS Abstract	Raven's SPM	MAT Accuracy	MAT DE	MAT SoW
Participant Occupational Prestige	.11	.19	.06	**.25***	**.35****	-.18	**.43****
Participant Educational Level	**.29****	**.20***	**.22***	**.41****	**.30****	.06	**.24***
Participant Social Status	.11	.21	.05	**.26***	**.33****	-.10	**.41****
Familial Occupational Prestige	.11	.11	.07	**.20***	.09	.09	.04
Familial Educational Level	.08	.15	.15	**.30****	.16	.15	.05
Familial Social Status	.11	.14	.09	**.25****	.12	.12	.05

Table 62. Matrix of Pearson correlations between measures of ability and socio-economic variables. Note: *$p < .05$; **$p < .01$

The pattern of correlations in Table 60 support H_3. Based on Cohen's effect sizes for correlation coefficients (Cohen, 1988), both conceptualisations of social status used in the study are weak-moderately positively correlated with all measures of ability in the study with the exception of the PfS Abstract. However, examining social status alongside the factors that contribute to it, performance on certain tests appears to be associated much more strongly with some socio-environmental variables than with others. Across all measures, test performance is most strongly related to participant educational level. This is, perhaps, unsurprising, as cognitive ability is known to predict academic achievement (Rohde & Thompson, 2007), meaning that one might expect higher mean test performance for those with more advanced qualifications. Examining the patterns of correlations for specific tests, it appears that the Raven's SPM is most strongly associated with all the socio-environmental factors, followed by the PfS Numerical, the PfS Verbal, and the PfS Abstract, this showing the least association with the other socio-environmental variables. Interestingly, it appears that performance on the PfS Numerical is more strongly associated with a participant's present socio-economic status than their familial one during development, whereas

performance on the PfS Verbal appears to be associated with approximately equal strength to both present and familial socio-economic status.

Though these results provide some insight into the socio-economic factors that are associated with performance on a range of tests, they cannot explain the mechanism by which performance differences arise. Specifically, there is no way of telling whether these socio-economic differences are reflective of true differences in ability, rather than differences in performance facilitating or debilitating factors. To do this, these factors must be unpacked to examine both their relationship to test performance and to their potential antecedents.

4.3.3.2 Test Familiarity as a Performance Facilitating Factor

As was demonstrated in section 4.2.1.1, mean scores on the test familiarity scale were moderately positively correlated with scores on all measures of ability used in the study (with the exception of the MAT's Speed of Working scale, to which it was unrelated). However, this alone does not give any indication of the extent to which higher test familiarity translates into better performance on these measures. To explore the practical effect that test familiarity might have on test performance, a series of simple regression models were run, in each of which total raw score was regressed on to mean test familiarity score. The B coefficients and β weights generated from these analyses (shown in Table 63) were then examined in an attempt to quantify for each test the number of raw score points associated with an increase in test familiarity.

	R^2	B	β
PfS Verbal	.08	1.21	0.28**
PfS Numerical	.08	1.61	0.29**
PfS Abstract	.06	3.41	0.25**
Raven's SPM	.07	1.78	0.27**
MAT Accuracy	.04	1.57	0.19*
MAT Decision Efficiency	.04	0.25	0.21*

Table 63. R^2, B coefficients and β weights for simple linear regressions of test familiarity predicting total test score. Note: *$p < .05$; **$p < .01$

Examining the β coefficients in the table above, it appears that an increase of one point in mean test familiarity is associated with fairly consistent increases of around 0.2 to 0.3 S.D. units in total raw score across all tools in the study. In terms of raw score points, a one-point increase in test familiarity appears to be associated with a 1- to 2-point step in mean raw score (the exceptions being for the PfS Abstract, for which there are many more items than the other measures, and MAT Decision Efficiency, which is a measure of rate of correct answers rather than raw score points *per se*). This supports H_4. These results might suggest that the construct measured by the test familiarity scale has a beneficial effect on test performance. However, due to the cross-sectional nature of the study, it would be dangerous to conclude that this causal relationship exists without further investigation into both its antecedents, and the mechanism by which it might influence performance.

To this end, a set of bivariate Pearson correlations were calculated between mean test familiarity scores and the socio-environmental variables examined in the previous analysis. The resultant correlations are shown in Table 64 below.

	Test Familiarity
Participant Occupational Prestige	.30**
Participant Educational Level	.42**
Participant Social Status	.39**
Familial Occupational Prestige	.33**
Familial Educational Level	.10
Familial Social Status	.30**

Table 64. Matrix of Pearson correlations between test familiarity and key variables in the dataset. Note: *$p < .05$; **$p < .01$

The correlations in Table 64 suggest that test familiarity is positively correlated with all of the socio-environmental variables measured in the study. Broadly speaking, this lends support to the exposure hypothesis of test familiarity development over time. Examining these variables more closely gives further insight into the specifics of how this development might take place. In the case of a participant's own socio-environmental context, both

occupational prestige and educational level appear to predict test familiarity. However, of these two factors, participant educational level is the much stronger predictor. Conversely, at the familial level, though parental educational background does seem to contribute weakly to development of test familiarity, familial occupational prestige appears to be much more strongly associated with it. Considering these findings together, they would appear to provide tentative support for the hypothesis that test familiarity is a construct that develops over one's lifetime, that is influenced not only by one's circumstances in childhood, but by subsequent academic and career progression. However, further analyses need to be conducted to fully understand how socio-environmental differences translate into test performance differences through the mechanism of increased test familiarity.

To further examine these relationships, a series of mediation analyses were conducted, to examine the circumstances under which test familiarity acts as a mediator between socio-economic variables and test performance. These analyses followed the logic of the process described by Baron and Kenny (1986) to test for mediation. However, rather than being based on a series of regression models, mediation was conducted in Mplus using SEM. In this approach, all paths are estimated simultaneously, as part of a single model. Mediation using SEM has been demonstrated to be consistently superior to the regression method, and allows mediation to be conducted on much smaller samples (Iacobucci, Saldanha & Deng, 2007). One drawback of this approach, however, is that it estimates the direct effect in the presence of the mediating variable, thus does not give any indication of the strength of the direct effect when the mediator is excluded from the model.

For this reason, it is important that mediation models are specified sensibly, and that a model is not specified in circumstances under which there is no substantive relationship between IV and DV. To this end, any cases in which any of the socio-environmental variables were demonstrated to be unrelated to a particular test score variable ($r < \pm.10$) were excluded from the analyses, as was the MAT Speed of Working variable (as it had been demonstrated to be unrelated to test familiarity). The remaining variables were all

standardised to *z*-scores, to allow comparison of direct and indirect effects. All effects were bootstrapped with 5000 resamples (based on the recommendation of Preacher & Hayes, 2004). Any effect (other than the direct effect) for which zero was observed to lie between the upper and lower bounds of the bootstrapped 95% confidence interval was excluded from further analyses. A summary of these analyses is shown in Table 65.

	IV → M	**IV → DV**	**M → DV**	**Indirect Effect**	**Mediation**
P. Occ. Prestige					
PfS Verbal	0.34**	(0.02)	0.30**	0.10*	Full
PfS Numerical	0.34**	(0.13)	0.23*	0.08	Full
Raven's SPM	0.34**	0.18	(0.15)	0.05	-
MAT Accuracy	0.34**	0.34**	(0.05)	(0.02)	-
MAT DE	0.34**	-0.27	0.21	0.07	Inconsistent
P. Edu. Level					
PfS Verbal	0.42***	0.19*	0.22*	0.10*	Partial: 33.9%
PfS Numerical	0.42***	(0.09)	0.26*	0.11*	Full
PfS Abstract	0.42***	(0.11)	0.24*	0.10*	Full
Raven's SPM	0.42***	0.35***	(0.12)	(0.05)	-
MAT Accuracy	0.42***	0.26**	(0.09)	(0.04)	-
P. Social Status					
PfS Verbal	0.43***	(-0.02)	0.31**	0.13*	Full
PfS Numerical	0.43***	(0.14)	0.21	0.09	Full
Raven's SPM	0.43***	(0.17)	(0.15)	0.06	-
MAT Accuracy	0.43***	0.31**	(0.06)	(0.03)	-
F. Occ. Prestige					
PfS Verbal	0.33***	(0.02)	0.28**	0.09*	Full
PfS Numerical	0.33***	(0.03)	0.27*	0.09*	Full
Raven's SPM	0.33***	(0.12)	0.23**	0.08*	Full
F. Edu. Level					
PfS Numerical	(0.10)	(0.12)	0.29**	(0.03)	-
PfS Abstract	(0.10)	(0.11)	0.24**	(0.02)	-
Raven's SPM	(0.10)	0.27**	0.24**	(0.02)	-
MAT Accuracy	(0.10)	(0.14)	0.18	(0.02)	-
MAT DE	(0.10)	(0.13)	0.19*	(0.02)	-
F. Social Status					
PfS Verbal	0.31**	(0.02)	0.28**	0.09*	Full
PfS Numerical	0.31**	(0.07)	0.26*	0.08*	Full
Raven's SPM	0.31**	0.18	0.22**	0.07*	Partial: 27.3%
MAT Accuracy	0.31**	(0.07)	0.18	0.06	Full
MAT DE	0.31**	(0.06)	0.20*	0.06	Full

Table 65. Analyses of test familiarity as a mediator of the relationship between socio-environmental variables and test performance. Note: *$p < .05$; **$p < .01$; ***$p < .001$; Coefficients in brackets denote that zero lies between upper and lower limits of the 95% bootstrapped confidence interval.

The analyses summarised in Table 65 provide further insight into the nature of test familiarity. It would appear that the test familiarity fully or partially mediates many of the relationships between the socio-environmental variables and test performance. The most consistent mediation effects were observed for the PfS Verbal and Numerical tests, for which test familiarity mediated the relationships between them and all the socio-environmental variables to some extent, with the exception of familial educational level. In addition to this, many other mediation effects of test familiarity were observed, particularly for the relationships between participant educational level, familial occupational prestige, and familial social status and a number of the ability measures used in the study. In particular, test familiarity appears to mediate the relationships between familial social status and performance on all measures of ability, with the exception of the PfS Abstract.

One particular mediation effect of test familiarity worthy of discussion was that observed for the relationship between participant occupational prestige and MAT Decision Efficiency. In this model, test familiarity is positively related to both participant occupational prestige and the MAT Decision Efficiency score, but participant occupational prestige is negatively related to Decision Efficiency. This produces an overall effect of the kind that MacKinnon, Fairchild and Fritz (2007) refer to as *inconsistent mediation*. In cases such as these, the mediator has a suppressing effect, appearing to increase the magnitude of the direct effect. There is no rational explanation for why participant occupational prestige is negatively related to MAT Decision Efficiency, particularly when positive effects have been observed between the majority of socio-environmental variables and test performance variables. It is most likely, then, that this result is a statistical artefact of the sample.

In spite of this anomaly, the findings of the mediation analyses appear to support H_5. These findings provide some tentative insight into test familiarity's effect on performance, and the mechanism by which it develops. These results suggest that the familial socio-economic environment in which a participant develops and the length of time that they remain in education both have an influence on the likelihood that they are exposed to testing of some

form. This increased exposure might lead them to develop increased test familiarity, which might, in turn, facilitate their performance on certain types of test. However, what is less clear at this stage is why this relationship appears to exist for some tests but not others. To truly understand these relationships, it is necessary to examine the proposed mechanism by which test familiarity might facilitate performance.

4.3.3.3 Explaining Differences across Tests in the Facilitative Effect of Test Familiarity

In order to understand the mechanism by which test familiarity affects performance, it was first necessary to investigate the degree to which each of the tests in the study loaded on general mental ability. The study's theoretical model suggests that test familiarity may influence test performance through the retention of schemata of the general forms of ability test items, effectively overlapping with *gf* as a construct in terms of the portion of fluid reasoning that relates to pattern recognition.

To investigate this, a CFA model was run on the data to look at *g*-loading of the tests in the study. Before this, however, it was deemed necessary to examine the patterns of intercorrelation between scales used in the study to identify those that would be most suitable for inclusion in this model. Table 66 below shows the matrix of Pearson correlations generated between all measures of ability used in the study.

	PfS Verbal	PfS Numerical	PfS Abstract	Raven's SPM	MAT Accuracy	MAT DE	MAT SoW
PfS Verbal	1						
PfS Numerical	.38**	1					
PfS Abstract	.56**	.32**	1				
Raven's SPM	.39**	.34**	.46**	1			
MAT Accuracy	.42**	.36**	.37**	.43**	1		
MAT DE	.37**	.33**	.25**	.26**	.45**	1	
MAT SoW	-.15	-.04	-.13	.02	.23**	-.60**	1

Table 66. Matrix of Pearson correlations between scores on ability measures. Note: *$p < .05$, **$p < .01$

Examining this correlation matrix, all of the measures appear highly correlated with one another with the exception of the MAT's Speed of Working scale. This scale appears to be largely orthogonal to every other non-MAT scale in the study, raising doubts about its potential to be used as a sensible measure of any apsect of cognition.

On the basis of these findings, it was judged most appropriate to select the PfS Verbal, PfS Numerical, PfS Abstract, Raven's SPM, and the MAT's Accuracy scale to be entered into the CFA model. While there would be nothing inherently wrong with including the MAT's Decision Efficiency scale in the model, its inclusion along with the Accuracy scale from the same tool (with which it is strongly correlated) might have introduced redundancies into the model, potentially skewing the factor loadings of the other scales. Observed variables to represent each of these five scales were entered into a CFA model and were made to load on to a single latent factor to represent *g*. The CFA model, including the standardised estimates of factor loadings for each indicator on to *g* is shown in Figure 26 below.

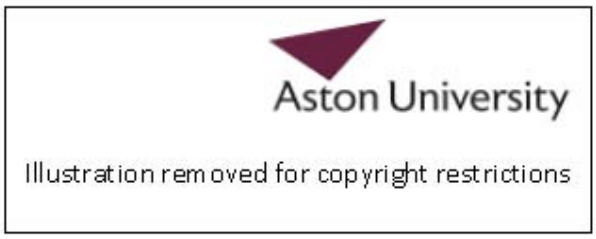

Figure 26. CFA model of ability measures made to load on a single latent factor, *g*.

The model was an excellent fit to the data (χ^2 = 6.84 *df* = 5; *p* = .23; CFI = .99; RMSEA = .05). Examining the factor loadings in Figure 26, it would, at first glance, appear that test familiarity facilitates performance most for tests that are highly *g*-loaded. The factor loadings on *g* for the PfS Verbal and the PfS Abstract are both very large at over .70, and both of these tests appear to be strongly facilitated by test familiarity. However, the factor loading for the PfS Numerical is substantially smaller than those of the other PfS tests at .52. Given that the PfS Numerical appears to be the test whose performance is most strongly associated with test familiarity, this seems inconsistent with H_6.

Furthermore, these factor loadings appear not to be proportional to the degree of variance that can be explained by ethnicity in each test observed in section 4.3.2. In particular, the

PfS Abstract's very strong *g*-loading appears inconsistent with the small amount of variance within it that can be explained by ethnicity. Therefore H_2 is not supported by these results.

On closer examination, there are some inconsistencies in the model with what one would expect based on the previous findings in the literature. Specifically, we would expect to observe stronger factor loadings for the MAT Accuracy scale (being a measure of WMC) and the Raven's SPM, given that these are both known to be very strongly *g*-loaded. It is possible, then, that the latent factor does not represent *g*, but, instead, a facet of it. For example, if the latent factor in the model were to refer to g_c rather than *g* itself, this might explain the inconsistent factor loadings in Figure 26. Verbal tests are known to tap heavily into g_c due to the crystallised knowledge that contributes to performance on these tests (through language skills, vocabulary, etc.), so this might explain the very high factor loading between PfS Verbal scores and the latent factor. However, the Raven's tools have been shown to have very little component attributable to g_c, so this line of reasoning does not really hold.

The implication of these findings is that the mechanism by which test familiarity facilitates test performance does not appear to depend on a test's *g*-loading. It is possible that test familiarity, instead, may have an effect that is non-cognitive in nature, depending less on pattern recognition and more on some aspect of how candidates approach tests in general. This would be supported by the finding that test familiarity has a facilitating effect on performance on the MAT. Since the rules necessary to correctly solve the problems presented in each trial are made explicit before each trial, test familiarity cannot affect performance through the mechanism of the schematic representation of reasoning test rules.

4.3.3.4 Personality Traits as Performance Facilitating and Debilitating Factors

To gain further insight into the nature of performance facilitation and debilitation, it is now necessary to consider the effects of some of the other factors proposed by the theoretical model. This section will consider, first, how personality traits relate to test performance, before considering how these traits interact with other key variables in the study.

Pearson correlations were calculated between raw scale scores on each of the 13 Trait scales and the cognitive measures in the study. The resultant correlation matrix is shown in Table 67 below.

	PfS Verbal	PfS Numerical	PfS Abstract	Raven's SPM	MAT Accuracy	MAT DE	MAT SoW
Achievement	.07	.11	-.16	.05	-.03	.06	-.03
Calmness	-.10	-.05	-.03	.06	.10	-.11	**.22***
Compassion	-.06	-.12	-.14	.01	.01	**-.22***	**.22***
Cooperation	-.02	-.07	-.06	.01	.04	-.17	**.19***
Culture	.05	-.05	.11	.17	.03	-.07	.02
Industriousness	-.10	-.07	-.14	-.02	-.04	-.13	.14
Intellect	**.32****	.17	.10	.18	**.20***	.01	.16
Leadership	-.09	-.00	**-.20***	-.12	-.05	-.11	**.22***
Optimism	**-.20***	-.05	**-.20***	-.02	.03	**-.21***	**.29****
Orderliness	-.05	-.07	-.13	-.13	-.15	-.09	-.07
Sensitivity	-.15	-.06	-.17	-.13	-.04	**-.19***	.18
Sociability	-.12	.02	-.17	-.09	-.07	-.09	.15
Stability	-.08	-.00	-.03	.10	.15	.07	.10

Table 67. Matrix of Pearson correlations between ability test scores and Trait personality scales. Note: *$p < .05$; **$p < .01$

The matrix of correlations between personality traits and ability test scores displayed in the table above revealed some unusual findings. As predicted, Trait scales related to Agreeableness (Compassion, Cooperation and Sensitivity) did not show significant correlations with any of the ability measures. This supports $H7_d$. However, of the remaining Traits, very few of the scales measuring them showed relationships with the cognitive measures that could reasonably be attributed to anything other than random error variance. The correlations between ability measures and Trait scales related to Conscientiousness (Orderliness and Industriousness) were all negative as predicted, though both the

significance level and practical effect size of these relationships were smaller than expected. This does not support H_{7c}.

There were, however, two exceptions to these observations. The first was that the Intellect scale (a facet of Openness) was correlated with scores on the PfS Verbal. As a high Intellect score indicates that a person displays preferences for abstract thinking and intellectual pursuits, it is perhaps unsurprising that scores on this scale are positively correlated with performance on a measure that likely depends heavily on crystallised intelligence. This finding supports H_{7b}. However, similar positive correlations can be observed between Intellect and all measures in the study. While these correlations are not significant at the .05 level, they represent weak correlations according to Cohen's effect sizes (Cohen, 1988).

The second exception was that there was a pattern of negative correlations between Optimism and most of the ability measures. While these correlations were fairly modest for many of the tools, they were pronounced for the PfS Verbal, the PfS Abstract and the MAT's Decision Efficiency scale. It is unclear why it should be the case that participants who score lower on the Optimism scale (a scale that has been shown to load most strongly on Extraversion and Emotional Stability) tend to score better on ability measures. This finding is particularly surprising, given that the Trait scales that directly measure facets of Extraversion (Sociability and Leadership) and Emotional Stability (Stability) do not appear to be related to ability in the way predicted in the study's hypotheses (and, therefore, do not support H_{7e} and H_{7a}, respectively).

The findings for Stability and Optimism could potentially be explained by the non-linear nature of their relationship with test performance. To examine this, linear regression solutions were compared to both quadratic and cubic solutions for these relationships. In the case of Stability, though a quadratic solution did represent an improvement in R^2 over the linear solution, none of these models were significant. This does not support H_{7f}. When

examining Optimism, neither the quadratic nor cubic solutions represented improvement in explained variance over the linear ones predicting test scores, with the exception of that predicting Raven's SPM scores (though these models were not significant). This does not support H_{7g}.

It would appear, then, that the only traits in the present study that showed a consistent relationship with test performance were Intellect and Optimism, though the latter of these was not related to test performance as predicted by the study's hypotheses. In order to properly understand how these two traits affected performance, and – thus – how they fit into the theoretical model, it was necessary to explore how they related to the other important variables in the study that have been linked to performance. Pearson correlations were calculated between the 13 Trait scales and these key variables. This correlation matrix is shown in Table 68 below.

	Test Fam.	P. Occ. Prestige	P. Edu. Level	P. Social Status	F. Occ. Prestige	F. Edu. Level	F. Social Status
Achievement	.19	.10	.15	.19	.12	**.23***	.18
Calmness	.04	.09	.17	.11	-.05	.06	-.03
Compassion	-.06	**.24***	.17	.19	.06	-.09	.02
Cooperation	-.02	**.29***	.13	**.25***	-.03	-.10	-.06
Culture	.05	.08	**.20***	.13	.16	**.22***	**.19***
Industriousness	.07	.25*	.13	.20	-.06	-.04	-.06
Intellect	**.21***	.10	**.36****	.17	.03	.16	.07
Leadership	**.21***	.14	.19	.22	.17	.07	.16
Optimism	.15	**.28***	.14	**.28***	.16	.11	.14
Orderliness	-.04	.01	-.20	-.12	-.19	**-.24***	**-.22***
Sensitivity	-.08	.16	.15	.13	.02	-.07	-.01
Sociability	**.21***	.13	.22*	.21	.12	-.00	.10
Stability	.10	.08	.22*	.15	-.02	.10	.01

Table 68. Matrix of Pearson correlations between personality traits and other key variables in the study. Note: *$p < .05$; **$p < .01$

Examination of the correlation matrix in Table 68 helps to crystallise understanding of some of the relationships between personality traits and ability test performance observed in the previous analysis. The Intellect scale is positively correlated with both test familiarity (r = .21) and participant educational level (r = .36). This suggests that high Intellect scores

predict the kinds of behaviours that are likely to increase test familiarity. Those who score more highly on the Intellect scale are more likely to have obtained higher educational qualifications, increasing their exposure to testing in the way described in section 4.3.3.2. It is also likely, however, that those with higher Intellect scores would have gained more exposure to ability testing than those who score low on this scale by way of their natural predisposition for intellectual pursuits.

In terms of the Optimism scale, its effect on test performance is somewhat more complex when contextualised by the findings in Table 68. Optimism is most strongly related to a participant's occupational prestige and their social status, though it appears to be positively correlated with all the socio-environmental variables in the study. This is consistent with the literature, in that positive correlations have previously been observed between participant SES and optimism, and between one's socio-economic status during development and optimism in adulthood (Ek, Remes & Sovio, 2003; Heinonen et al., 2006). This might suggest Optimism as a possible mediator of the relationship between socio-economic status and test performance. However, Optimism was observed to be negatively correlated with performance on many measures of ability, ruling out this mediating relationship. It is possible that this negative relationship is a statistical artefact. To examine this, it will be necessary to examine the effect of Optimism on test performance having controlled for the effect of ethnicity.

4.3.5 Ethnic Differences on Key Variables

Having explored the nature of performance facilitation in the dataset, the final stage of the analyses was to investigate the degree to which ethnic differences in test performance could be accounted for by differences in the levels of these performance facilitating and debilitating factors. Before doing this, it was first necessary to investigate the degree to which these factors varied across ethnic groups.

Means, standard deviations, and values of Cohen's *d* for each of the key variables in the study were calculated for each ethnic group according to the two conceptualisations of ethnicity. These results are shown below in Tables 69 and 70.

Ethnic Group		**White**	**Non-White**
Occupational Prestige			
	N	58	24
	Mean	5.05	4.29
	(S.D.)	(3.32)	(3.36)
	Cohen's *d*	-	0.23
Educational Level			
	N	71	41
	Mean	4.83	4.32
	(S.D.)	(1.62)	(1.75)
	Cohen's *d*	-	0.30
Social Status			
	N	58	21
	Mean	40.52	40.43
	(S.D.)	(19.60)	(19.11)
	Cohen's *d*	-	0.01
Familial Educational Level			
	N	70	49
	Mean	3.92	3.55
	(S.D.)	(1.68)	(1.84)
	Cohen's *d*	-	0.21
Familial Occupational Prestige			
	N	71	51
	Mean	5.60	3.82
	(S.D.)	(2.07)	(3.06)
	Cohen's *d*	-	0.68
Familial Social Status			
	N	70	49
	Mean	39.85	29.95
	(S.D.)	(13.91)	(18.16)
	Cohen's *d*	-	0.61
Test Familiarity			
	N	71	51
	Mean	3.36	3.03
	(S.D.)	(0.94)	(0.79)
	Cohen's *d*	-	0.38
Intellect			
	N	64	48
	Mean	3.97	3.81
	(S.D.)	(0.59)	(0.90)
	Cohen's *d*	-	0.21
Optimism			
	N	64	48
	Mean	3.81	3.66
	(S.D.)	(0.90)	(0.70)
	Cohen's *d*	-	0.19

Table 69. Mean raw scores and Cohen's *d* on key variables by ethnicity classified as White/Non-white.

Ethnic Group		White	Black	Asian
Occupational Prestige				
	N	52	13	11
	Mean	5.10	2.85	6.33
	(S.D.)	(3.24)	(2.85)	(2.65)
	Cohen's *d*	-	0.74	-0.42
Educational Level				
	N	62	13	25
	Mean	4.79	4.31	4.36
	(S.D.)	(1.57)	(1.44)	(1.93)
	Cohen's *d*	-	0.32	0.24
Social Status				
	N	52	11	10
	Mean	40.60	30.00	50.67
	(S.D.)	(19.13)	(16.81)	(15.13)
	Cohen's *d*	-	0.59	-0.58
Familial Occupational Prestige				
	N	61	15	30
	Mean	5.58	4.30	3.37
	(S.D.)	(2.13)	(2.84)	(2.98)
	Cohen's *d*	-	0.51	0.85
Familial Educational Level				
	N	62	15	32
	Mean	3.80	4.37	3.02
	(S.D.)	(1.72)	(1.65)	(1.66)
	Cohen's *d*	-	-0.34	0.46
Familial Social Status				
	N	61	15	30
	Mean	39.39	34.60	26.13
	(S.D.)	(14.33)	(15.28)	(17.71)
	Cohen's *d*	-	0.32	0.82
Test Familiarity				
	N	62	15	32
	Mean	3.34	3.03	3.03
	(S.D.)	(0.94)	(0.89)	(0.73)
	Cohen's *d*	-	0.33	0.36
Intellect				
	N	56	13	30
	Mean	3.95	3.96	3.77
	(S.D.)	(0.61)	(0.69)	(0.53)
	Cohen's *d*	-	-0.01	0.24
Optimism				
	N	56	13	30
	Mean	3.74	4.05	3.57
	(S.D.)	(0.91)	(0.60)	(0.69)
	Cohen's *d*	-	-0.40	0.21

Table 70. Mean raw scores and Cohen's *d* on key variables by broad ethnic group.

Examining Tables 69 and 70, the patterns of ethnic group differences on these key variables mirror those for the ability test raw scores in section 4.3.1 closely. In particular, there are substantial differences between ethnic groups in terms of their educational level, and these differences equate broadly to differences in mean test familiarity scores between these

groups. This suggests that test familiarity – as a performance facilitating factor – may be able to account for some of the differences in raw test score between ethnic groups for some of the measures in the study.

When examining ethnic group differences in Optimism scale scores, the scale does not appear to behave as expected based on previous analysis. Specifically, in the breakdown of ethnicity by broad ethnic group, the Black group showed a substantially lower level of both occupational prestige and participant social status relative to the White majority group, whereas the Asian group demonstrated substantially higher levels on these variables. However, when considering these groups' mean Optimism score, the reverse appears to be the case (i.e. the Black group scored substantially higher and the Asian group substantially lower on Optimism than the White majority).

One possible explanation for this is that the relationship between optimism and socio-economic variables is not stable across ethnic groups. This may, at first, sound far-fetched, but it is not without precedent in the literature. Though it has been observed that optimism is positively correlated with SES for White, Western samples (e.g. Ek, Remes & Sovio, 2003; Heinonen et al., 2006), no correlation was found between these two variables in a sample of Colombian managers (Juárez & Contreras, 2012). Differences in the nature of relationships have also been observed between ethnic groups within the US. Schutte, Valerio and Carrillo (1996) observed a much smaller (and not statistically significant) correlation between optimism and SES in a sample of Mexican-Americans when compared to that for a sample of Anglo-Americans, attributing these differences to factors associated with the primarily collectivist culture of the former ethnic group, and the primarily individualist culture of the latter. African and Indian Subcontinent cultures are known to be primarily collectivistic (Hofstede, 1980), so it is possible that the level of trait optimism for people whose ethnic origin derives from these cultures is less tied to SES than it is for the White majority group. In the present study's sample, it may be the case that Optimism is a performance debilitating factor (given its negative correlation with performance) that is responsible for a degree of the

ethnic group performance differences that have been observed. If this were the case, as opposed to having its root cause in socio-economic factors, this factor would likely be rooted in cultural factors, derived from internalised familial norms.

It is possible that these two variables can account for a proportion of the variance in test scores that can be explained by ethnicity. To investigate this in a more robust way, the influence of ethnicity on test performance was examined when controlling for these variables, individually at first, and then when combined. The remaining sections of this chapter will examine how able test familiarity and trait optimism are to account for ethnic group performance differences in the sample.

4.3.6 Test Familiarity Differences as an Explanation for Group Differences

First, test familiarity was examined as a performance facilitating factor that might potentially explain a degree of the observed ethnic group difference in test scores. In order to do this, a series of hierarchical regression models predicting ability test raw scores were run. In each analysis, test familiarity score was entered into the regression model in Step 1. In Step 2, dummy coded variables were added to the model to represent ethnicity. The change in magnitude of multiple *R* in Step 2 was recorded. This was then compared to the values of *R* observed between test scores and ethnicity in section 4.3.2 (i.e. without the control of test familiarity). Multiple *R* was chosen over R^2 in these analyses as R^2 has the potential to cloud observations when used to compare effect sizes, due to the exponential way in which it increases. A reduction in the size of multiple *R* with the inclusion of test familiarity in the regression models would indicate that it can account for at least some of the observed ethnic group differences in test performance.

These analyses generated a large amount of results. In order to simplify the findings of these analyses and allow them to be more easily interpreted, they were collated into Tables 71 and 72 below. These tables compare the multiple *R* between test scores and ethnicity

both with and without the controlling influence of test familiarity. To further contextualise these results, models for which there is a significant F change with the addition of ethnicity variables are flagged. It is worth noting for these tables (and for those in the subsequent analyses in this chapter), that some of their calculations may look inaccurate. Multiple *R* values in these tables are rounded to 2 decimal places. However, the differences in multiple *R* with and without control – and these differences expressed as a proportion of the total effect of ethnicity on test scores – are calculated using absolute values.

	Multiple *R* between test score and Ethnicity (White-Non-white)		
	Without Control (%)	**With Control (%)**	**Difference (as proportion of effect)**
PfS Verbal	.41***	.17***	.23 (57.1%)
PfS Numerical	.19*	.06*	.13 (67.5%)
PfS Abstract	.14	.02	.13 (87.4%)
Raven's SPM	.29**	.11**	.18 (62.1%)
MAT Accuracy	.33***	.18**	.15 (46.1%)
MAT DE	.17	.04	.13 (76.6%)

Table 71. Summary of the Multiple *R* between mean raw score and ethnicity (White-Non-white classification) that can be explained by test familiarity. Note: *indicates model is significant at .05 level; **indicates model is significant at .01 level; ***indicates model is significant at .001 level.

	Multiple *R* between test score and Ethnicity (Broad)		
	Without Control	**With Control**	**Difference (as proportion of effect)**
PfS Verbal	.45***	.23***	.22 (49.2%)
PfS Numerical	.24*	.10	.14 (60.2%)
PfS Abstract	.24*	.08	.16 (65.6%)
Raven's SPM	.26*	.10*	.16 (60.3%)
MAT Accuracy	.34**	.21**	.13 (39.0%)
MAT DE	.26*	.14*	.12 (46.1%)

Table 72. Summary of the Multiple *R* between mean raw score and ethnicity (broad classification) that can be explained by test familiarity. Note: *indicates model is significant at .05 level; **indicates model is significant at .01 level; ***indicates model is significant at .001 level.

Examining the preceding two tables, it would appear that, across all measures of ability, test familiarity can account for a substantial proportion of ethnic group test score differences. In terms of effect size, across all tests, it would appear that between 39% and 87% of the variance in performance explained by ethnicity can be accounted for by test familiarity. Given that these differences are being explained by a single variable, this is a huge amount of the effect between ethnicity and test scores. This would appear to support H_{10}.

Furthermore, this effect appears to be more pronounced for some tests than for others. Though findings are somewhat variable across conceptualisations of ethnicity, it seems that test familiarity is able to account, proportionally, for much more of the effect of ethnicity on raw scores for some tests than for others, for which there is much less overlap. Of the four measures of ability used in the study, ethnic test score differences on the PfS Numerical, the PfS Abstract, and the Raven's SPM appear to be more accountable by test familiarity than those for the PfS Verbal. However, this does not provide full support for H_{10}, in that these differences appear not to depend upon the proportion of a test's performance that depends upon g_f.

Nevertheless, this is an excellent first step in explaining ethnic group test performance differences in terms of non-cognitive factors. However, it still appears, in many cases, that there is more variance in test performance between ethnic groups to explain. The next section will consider Optimism as a potential performance debilitating factor.

4.3.7 Optimism as an Explanation of Ethnic Group Test Performance Differences

To examine the potential of trait Optimism to explain ethnic group test performance differences, a similar procedure was employed to that of the analyses in section 4.3.6. A series of hierarchical regressions were run to examine the multiple *R* between test scores and ethnicity, once Optimism scores had been controlled for. However, unlike test familiarity, which has been demonstrated to be positively associated with test scores on all

measures in the study, there were a number of measures to which Optimism had been demonstrated in section 4.3.3.4 to be unrelated (i.e. the PfS Numerical, the Raven's SPM and the MAT Accuracy). These findings were excluded from the analyses to avoid any anomalous findings due to poorer model fit in the hierarchical models for these measures over their simple regression models. A summary of the findings of these analyses are shown in Tables 73 and 74 below.

	Multiple *R* between test score and Ethnicity (White-Non-white)		
	Without Control (%)	**With Control (%)**	**Difference (as proportion of effect)**
PfS Verbal	.41***	.24***	.16 (40.4%)
PfS Abstract	.14	.07	.07 (48.3%)
MAT DE	.17	.04	.13 (76.6%)

Table 73. Summary of the Multiple *R* between mean raw score and ethnicity (White-Non-white classification) that can be explained by Optimism. Note: *indicates model is significant at .05 level; **indicates model is significant at .01 level; ***indicates model is significant at .001 level.

	Multiple *R* between test score and Ethnicity (Broad)		
	Without Control (%)	**With Control (%)**	**Difference (as proportion of effect)**
PfS Verbal	.45***	.18***	.27 (59.6%)
PfS Abstract	.24*	.07	.17 (69.3%)
MAT DE	.26*	.08	.18 (70.3%)

Table 74. Summary of the Multiple *R* between mean raw score and ethnicity (broad classification) that can be explained by Optimism. Note: *indicates model is significant at .05 level; **indicates model is significant at .01 level; ***indicates model is significant at .001 level.

Examining the two preceding tables, it would appear that, like Test Familiarity, Optimism can account for a large proportion of ethnic group test score differences across these three measures. Though the magnitude of effect size that can be explained by Optimism for these measures is somewhat smaller than those explained by Test Familiarity in the previous analyses, these still represent between 40 and 70% of the overall effect. This suggests that

Optimism – like Test Familiarity – could be an important non-cognitive factor in explaining ethnic group test score differences, particularly when ethnicity is broken down to broad ethnic classifications as opposed to being conceptualised as a White/Non-white dichotomy. Examining the findings for the PfS Abstract and MAT's Decision Efficiency scale, the introduction of optimism as a control reduces the significance of the F change due to the ethnicity variables above the .05 level. These findings support H_{11}. As was the case for test familiarity, though, there does appear to be a degree of variation in the proportion of the variance across tests that optimism scores can explain. At this stage, it is not entirely clear why this should be the case, but, again, it does appear that these effects do not depend on a measure's *g*-loading.

4.3.8 The Combined Effect of Performance Facilitating and Debilitating factors

The final step in the analyses was to establish whether, combined, these two non-cognitive factors could explain ethnic group test score differences in the measures to which they had previously been demonstrated to be associated. Given that there is little conceptual overlap between optimism and test familiarity as constructs, coupled with the weak-moderate correlation observed between the two (r = .15; p = .11), it was reasoned that these variables likely did not share excessive variance in explaining test scores. For that reason, in examining the multiple *R* between test scores and these two non-cognitive factors jointly, both were entered simultaneously in Step 1 in a series of hierarchical regression models. As before, ethnicity variables were entered in Step 2, and change in multiple *R* (and the significance of the F change of this step) compared to that of ethnicity without control. Summary tables for these analyses are shown below in Tables 75 and 76.

	Multiple *R* between test score and Ethnicity (White-Non-white)			
	Without Control (%)	**With Test Familiarity Control (%)**	**With Test Familiarity and Personality Controls (%)**	**Difference (as proportion of effect)**
PfS Verbal	.41***	.17***	.15***	.29 (62.8%)
PfS Abstract	.14	.02	.05	.09 (62.9%)
MAT DE	.17	.04	.04	.13 (74.5%)

Table 75. Summary of the Multiple *R* between mean raw score and ethnicity (White-Non-white classification) that can be explained by test performance facilitating and debilitating factors. Note: *indicates model is significant at .05 level; **indicates model is significant at .01 level; ***indicates model is significant at .001 level.

	Multiple *R* between test score and Ethnicity (Broad)			
	Without Control (%)	**With Test Familiarity Control (%)**	**With Test Familiarity and Optimism Controls (%)**	**Difference (as proportion of effect)**
PfS Verbal	.45***	.23***	.15***	.29 (66.1%)
PfS Abstract	.24*	.08*	.05	.19 (78.0%)
MAT DE	.26*	.14*	.04	.21 (83.6%)

Table 76. Summary of the Multiple *R* between mean raw score and ethnicity (broad classification) that can be explained by test performance facilitating and debilitating factors. Note: *indicates model is significant at .05 level; **indicates model is significant at .01 level; ***indicates model is significant at .001 level.

Examining Tables 75 and 76, it would appear that the combination of test familiarity and Optimism has a marked effect on the proportion of ethnic group differences that can be accounted for. While, for White-Non-white differences, the effect of adding Optimism to the model is only relatively modest when compared to that of test familiarity alone, the observed performance differences between broad ethnic classifications are much more readily explained when both predictors are considered side by side than when they are treated in isolation. The largest effect appears to be on ethnic differences in raw scores on the MAT's Decision Efficiency scale (83.6% of the total effect), though, to a degree, the proportion of broad ethnic group differences that can be explained benefits from the inclusion of both of these variables in the model as separate constructs. This supports H_{12}.

These results represent a huge step towards closing a knowledge gap that has dogged the ethnic test performance literature for a very long time. Seemingly, these two non-cognitive variables can account for a much larger proportion of ethnic group test score differences than has previously been possible. Given that both test familiarity and optimism are both theoretically orthogonal to ability, their combined efficacy in explaining ethnic performance differences provides great insight into the process by which these differences arise at the group level.

CHAPTER 5: GENERAL DISCUSSION

5.1 Chapter Overview

There were two overriding findings from the present research. The first was that Differential Test Functioning appears unable to explain more than a trivial proportion of ethnic group test performance differences. The second was that a very large proportion of these differences in performance on a variety of measures of ability can be explained by group differences on two non-cognitive factors: Test familiarity and optimism. Previously, neither of these factors has been properly examined as a plausible explanation for ethnic group performance differences. The present research sheds some much-needed light on the mechanism by which key socio-environmental factors give rise to differences in ability test performance that is not dependent on differences in the constructs assessed by these measures.

In this final chapter, the findings of both studies in the present research will be synthesised and their key findings extracted. These findings will be interpreted in the context of the extant ethnic bias literature, before their overall contribution to knowledge is explored. These contributions will then be formalised with the presentation of a revised theoretical model for the study, providing a framework for academics to further explore these observed phenomena in greater depth. Once the study's theoretical contribution has been explored, its practical impact will be considered. Recommendations will be made for how best to address ethnic group test performance differences at both an organisational level and at a societal one. Finally, the findings of the present research will be tempered by an analysis of its limitations. The chapter will conclude with an exploration of some possible avenues for future research, during which the author will call for a line of research to be initiated in order for some of these issues to be properly understood.

5.2 Summary of Key Findings

Interpreting the findings of the analyses together, light can be shed on a number of issues that have classically not been well understood in the literature. Combined, the two studies produced many and varied findings, and – as such – a degree of structure is required to be imposed upon them in order to interpret the impact that they have on our understanding of issues related to ethnic differences in test performance. The following sections will summarise the findings of both studies relevant to a number of key areas of the literature, examining where areas of consistency lie and, where inconsistencies can be identified, offering possible explanations for why these might have been observed.

5.2.1 Ethnic Group Test Score Differences

In Study 1, consistent ethnic group test performance differences were observed between the White majority group and many ethnic minority groups. In the SPM data archive, the majority of differences in raw score points between the White majority group and minority ethnic groups fell between 0.3 and 0.9 S.D. units (though, in a number of cases, these differences were smaller than this or non-existent, particularly for some of the specific ethnic groups in the dataset). Ethnic differences in the APM archive tended to be less pronounced, but still, in places, amounted to around 0.5 S.D. units. Across both datasets, the 'Other White' groups substantially outperformed the White British majority, as did the Chinese subgroup in the APM dataset.

The findings of the first study were supported by those of the second study. Broadly similar patterns of difference were observed between minority ethnic groups and the White (or White British) groups. In the majority of cases, persistent ethnic differences favouring the White majority were observed of between 0.3 and around 0.9 S.D. units. Once again, the non-British White group tended to score more highly than did the White British group.

All of these findings lend support to the first hypothesis in both studies that ethnic group differences exist across all measures of ability. However, with the exception of those for the PfS Verbal, these differences were of a magnitude smaller than those previously observed in the literature. Overall, between non-White ethnic groups and the White majority, group differences were smaller than those of 0.83 S.D. units observed by Schmitt, Clause and Pulakos (1996), and, in many cases, smaller than the 0.46 S.D. units difference observed by Martocchio and Whitener (1992). Additionally, the White-Non-white differences on the PfS measures observed in Study 2 were smaller for the PfS Verbal and PfS Abstract than Childs et al. (2013) recorded during validation of their tools, the difference for the PfS Numerical only being marginally larger. This adds further support to previous research (e.g. Dickens & Flynn, 2005; Woods, Hardy & Guillaume, working paper) that ethnic minority groups appear to be narrowing the gap on the White majority, albeit slowly. As has previously suggested by Dickens and Flynn (2005), this implies that the argument that ethnic differences are based primarily on differences between ethnic groups at the genetic level is difficult to sustain. The differences, then, are more likely to arise principally from differences in the opportunities afforded to different ethnic groups within society. The present research offered support for this position in the form of findings in two areas. Firstly, in Study 2, socio-economic variables at both the participant and familial levels status appeared to be weak-moderately positively correlated with test performance across all measures of ability with the exception of the PfS Abstract. These relationships were most consistently observed between ability test scores and participant educational level, the weakest of these being with scores on the MAT's Decision Efficiency scale (r = .06) and the strongest being with the Raven's SPM (r = .41). Secondly, consistent patterns of difference in social status were observed between ethnic groups, consistently favouring the higher performing White majority group. Again, these effects were most noticeable when social status was examined at the familial level. This implies that, at the group level, social status and variables related to it can, when examined at both the familial and current levels, explain a degree of ethnic difference in test performance.

However, there was a degree of inconsistency to the present study's findings relating to ethnic group test performance differences. Examining the patterns of ethnic group performance, differences between ethnic groups appear to be most pronounced for the PfS Verbal. At first glance, this seems consistent with the Spearman-Jensen Effect (Reeve & Bonaccio, 2009), that ethnic group differences increase as a measure's *g*-loading increases. However, the PfS Abstract was also shown to be comparably *g*-loaded to the PfS Verbal, but displayed ethnic group differences of much smaller magnitude in all cases. The fact that the Spearman-Jensen Effect could not be replicated in the present study suggests that group differences in *g* cannot adequately explain differences in performance between ethnic groups.

The Spearman-Jensen Effect (or a phenomenon related to it) has been observed for other tools, particularly in situational judgement tests (SJTs). SJTs tend to vary most strikingly in their *g*-loading dependent on the nature of the response instructions provided to candidates. Behaviour-based SJTs, in which candidates are asked how they *would* react to a given situation, tend to correlate much less strongly with measures of *g* than do knowledge-based SJTs, in which candidates are asked how they *should* react (McDaniel et al., 2007). Meta-analytic evidence provided by Whetzel, McDaniel and Nguyen (2008) has shown that ethnic group differences are far greater for knowledge-based SJTs than they are for behavioural SJTs. Whetzel, McDaniel and Nguyen argue that this observation is indicative of true differences in *g* between these groups. The findings of the present study, however, would not appear to support this hypothesis.

Having established that ethnic group test performance differences do still exist across a variety of ability measures, the present research sought to explain how the mechanisms by which these differences arise.

5.2.2 Explanations for these Differences

As discussed in Chapter 2, what causes some ethnic groups to consistently outperform others on measures of cognitive ability is still a point of fierce debate in the literature. Though many possible explanations for why ethnic groups might show relatively consistent differences in test scores have been proposed (e.g. Hough, Oswald & Ployhart, 2001), no real consensus has been reached as to why this might be the case. This represents an ongoing and severe knowledge gap in the literature. The present research's main aim was to address this gap, by re-examining two key areas of disagreement. First, in Study 1, the claim that ethnic differences could be accounted for by differential test functioning was examined. In Study 2, the root causes of these differences, and the mechanisms by which they might lead to increased or decreased performance, were explored. The findings of both studies relevant to these factors are summarised in the following sections.

5.2.2.1 Differential Test Functioning

Based on the review of the literature in Chapter 2, a likely source of ethnic group test performance differences was posited to be ethnic differential test functioning (DTF). Ethnic DTF has long been maintained to be a potentially important factor in explaining why some ethnic groups appear to perform less well on ability tests than others, as it seeks to isolate the variance in test performance that cannot be accounted for by differences in the construct of interest (i.e. *g*). However, previous research in this area has been criticised in that the majority of published ethnic DIF/DTF studies suffer from three serious methodological problems (Hunter & Schmidt, 2000): violations of the assumption of unidimensionality, reliance on significance testing, and failure to control for measurement error in ability estimates. Aside from these issues, Stark, Chernyshenko and Drasgow (2004) have argued that, in cases in which ethnic DIF effects can be detected, they tend only to contribute to a trivial overall DTF effect. In their study, they observed differences between ethnic groups

attributable to DTF of between 0.02 and 0.25 S.D. units. These effects, they argue, are not sufficient to explain a substantial proportion of ethnic group differences.

Study 1 set out to clarify the literature on the effects of ethnic DTF by using a more robust approach to its identification than has typically been employed in previous studies, addressing the concerns raised by Hunter and Schmidt (2000). It did so using two large, new, unpublished datasets. While the study used a more traditional method for the identification of ethnic DIF (and DTF) in the form of the LR Method, it also made use of MLVM, a more sophisticated approach to the investigation of DTF, based on the observation of Cohen and Bolt (2005) that DIF/DTF identification on the basis of manifest characteristics (such as ethnicity) is rarely fruitful. The results of Study 1 revealed that, for both the SPM and the APM, neither approach conclusively identified the existence of ethnic DTF. Though the MLVM analyses did suggest that more than one latent response class existed in each of the samples, EFA conducted to confirm the unifactorial underlying structure of both tools cast some doubts on the validity of these observations in that, for both the SPM and APM, two substantive factors appeared to underlie the items. The first of these factors appeared to be most strongly associated with the earlier items of each test, whereas the second was associated with the later items. To corroborate this, when the nature of the MLVM latent classes were explored, it appeared that the classes differed most in terms of the items for which they had the greater odds of a correct response, members of one latent class demonstrating an advantage for the earlier items, and another for the later items.

Hunter and Schmidt (2000) have cited violations of the assumption of unidimensionality to be among the most frequent issues with DIF/DTF studies. Furthermore, Bauer and Curran (2004) suggest that – in the specific case of mixture models – violation of this assumption can lead to misspecification of the model, increasing the likelihood of spurious latent classes. It would appear, then, that this might have been the case for the MLVMs conducted in the present study. To explore this further, in Study 1 a follow-up MLVM was conducted using an approach that is rarely employed in the literature, with the intention of clarifying the results of

the previous MLVM. This approach split the Raven's measures into subtests made up only of items that loaded meaningfully on to a single factor underlying them. The effect of this was to make each of the simplified mixture models unifactorial, meeting the assumptions of MLVM. However, when MLVM was run on these simplified models, no evidence of sample heterogeneity could be identified, suggesting that the latent classes identified in the full MLVM analysis were most likely spurious, based on differences in factor scores between the two factors underlying each test.

These results suggest that DTF as an explanation for substantial ethnic group test score differences is – at least in the case of the Raven's measures – unsatisfactory. These findings may go some way to resolving a debate that has been ongoing for some time. Rationally, it makes a degree of sense that cultural differences might lead test takers of different ethnicities to interpret – and, therefore, respond to – a test item in different ways (e.g. Hough, Oswald & Ployhart, 2001; Ibarra, 2001; Li, Cohen & Ibarra, 2004). However, this argument is frequently not borne out by empirical findings as a way of explaining meaningful differences in ethnic group test performance (Stark, Chernyshenko & Drasgow, 2004). Based on these findings, it appears that much of the research evidence that claims that a meaningful proportion of ethnic differences can be explained by DTF may need to be carefully reassessed.

5.2.2.2 Test Familiarity as a Performance Facilitating Factor

In Study 2, a number of potential performance facilitating and debilitating factors were examined as possible explanations for ethnic group performance differences. The first of these to be considered was test familiarity. Though a degree of research attention has been paid in the past to test familiarity's effect on performance (e.g. DerSimonian & Laird, 1983) – and even as a potential focal point by which to reduce ethnic group test score differences (e.g. Frierson, 1986; Sackett et al., 2001) – the present research argued that what the

literature lacked was a clear working definition that put clear blue water between itself and test-specific, ephemeral concepts such as practice effects and test-specific coaching, or those that conflated test performance with performance in academic examinations. All of these can be viewed as substantial knowledge gaps within the existing literature relating to what is potentially an extremely important construct.

Having developed a short and robust scale to assess participant test familiarity, Study 2 set out to assess its effect on test performance. The regression models in section 4.3.3.2 suggested that a one-point step increase in mean test familiarity was associated with score increases on all the measures of ability in the study of between 1.2 and 3.4 raw score points. The beta weights that were calculated in these analyses indicated that a mean increase of one point on the test familiarity scale was associated with an increase in test performance of between 0.2 and 0.3 S.D. units for all measures used in the study. These values are substantially larger than the effects typically noted in the educational literature for increases due to test coaching programmes.

Though this finding in and of itself is interesting and of some practical importance, this study's later findings revealed that the real value of test familiarity was in explaining ethnic group test score differences. It was demonstrated that ethnic group test scores differences can be substantially reduced – by between 39 and 87%, depending on the test and the conceptualisation of ethnicity used – by taking the mean test familiarity of ethnic groups into account. This suggests that the conclusion reached by Sackett et al. (2001) that increases in test familiarity do little to reduce ethnic group test performance differences may require re-examination. The findings of the present research would need to be confirmed in a follow-up intervention study, but these results are encouraging nonetheless.

In an attempt to explain how ethnic group differences in test familiarity might arise, the present research postulated an explanation based on exposure, the roots of ethnic group differences being in socio-economic differences between these groups. The present study

confirmed that, indeed, test familiarity appeared to be strongly related to differences on key socio-economic indicators during both development and adulthood, the strongest predictor of which was participant educational level. Furthermore, test familiarity was found to display a strong mediating effect on the relationship between almost all of the socio-environmental variables (with the exception of familial educational level) and test performance on the PfS Verbal and Numerical measures, and was found to mediate the relationship between familial social status and performance on all measures in the study apart from the PfS Abstract. While the use of the Social Status scale and its constituent elements as a conceptualisation of one's socio-economic status was never intended to be a definitive measure of the intricacies of each individual's developmental environment, the relationship between these factors nevertheless provides compelling evidence for the exposure hypothesis, that those from higher socio-economic backgrounds were more likely to be more familiar with testing, having encountered them more frequently throughout their lives.

In spite of this, test familiarity did not appear to behave in the way predicted by the study's theoretical explanation. It was reasoned that this increased test familiarity would facilitate test performance through increased exposure to general test forms, aiding pattern recognition, thus showing the greatest benefits for performance on measures that included a strong g_f component. However, when each test's *g*-loading was estimated using CFA, this did not appear to be the case. Furthermore, if this explanation were the case, one would expect test familiarity to be largely orthogonal to scores on the MAT scales. Given that instructions for each trial in this measure are made explicit before that trial commences, it seems unlikely that pattern recognition of forms and rules of test items would be associated with any facilitative effect on performance on these test items. However, test familiarity scores were shown to be positively correlated with scores on the MAT Accuracy and MAT Decision Efficiency scales (albeit less strongly than for the other ability measures). These observations appear to disconfirm the hypothesis that test familiarity aids test performance through exposure to the general forms of reasoning represented in specific test items. Other

possibilities for the mechanism by which test familiarity aids test performance would, therefore, have to be explored.

One possibility presents itself in the form of how test takers approach tests as a whole. Ability tests – and, in particular, those with a strict time limit ('speed tests') – require candidates to strike the correct balance between speed and accuracy when completing them. All ability tests have – by their nature – a multiple choice response format, and there is rarely any kind of penalty for choosing an incorrect answer. Furthermore, it is most frequently the case that all items confer the same score increase for a correct answer, so the first, easiest item is worth exactly the same amount as the last (theoretically) most difficult item. Therefore, it is in a candidate's interest to make an educated – or even a wild – guess if they cannot work out the correct answer to an item. This, however, is in direct opposition to the logic of academic examinations, in which wild guesses rarely pay off, and could even, potentially, damage one's overall mark by revealing a lack of knowledge in a particular area. Candidates, then, who are unfamiliar with ability testing are far more likely to approach them in the manner they would an academic exam, taking a slow, methodical approach to the items. If this were the case, these candidates might run the risk of dwelling too long on an early question that they found complex, devoting more of the limited time they had for the test than was appropriate. Conversely, one who was familiar with ability testing would know that time is extremely pressurised in these circumstances, and might even have gone to the lengths of calculating an approximate mean time per item before starting the test. This might lead candidates of this kind to be more aware than neophyte test takers of when they might be spending too much time on a single item, leading them to cut their losses and move on to the next item in the test.

The results of the present study provide a degree of support for this explanation, based on the patterns of relationships between test familiarity and the constructs assessed by the MAT. Scores on the test familiarity scale correlate moderately strongly with the MAT's Accuracy scale ($r = .19$; $p < .05$), but are much less strongly related to its Speed of Working

scale (r = -.07, p = .47). The strongest relationship test familiarity shows with the MAT subscales is that between it and the Decision Efficiency scale (r = .21; p < .05). Overall, these findings suggest that the more familiar a candidate is with ability testing in general, the more likely they are to find the most efficient balance between speed and accuracy when taking these tests. However, some of the findings of the present study appear inconsistent with this account. The PfS Abstract (and, to a lesser extent, the PfS Verbal and PfS Numerical) have a relatively large number of questions to be answered in a relatively short period of time. An efficient balance of speed and accuracy is far more important for these measures than it is for the Raven's SPM (which is considered to be a power test, and, as such, far less time-pressured). However, test familiarity appears to aid performance on all of these measures relatively consistently, showing, in actual fact, the smallest effect for performance on the PfS Abstract. Further exploration of the nature of test familiarity's observed performance facilitating effect will need to be conducted in the future to be able to clarify these relationships and to properly investigate this proposed mechanism.

5.2.2.3 Personality Traits and their Influence on Performance

While these findings are compelling, test familiarity was not able to explain ethnic group differences *in toto*. To build on these findings, the effect of differences in personality traits on the test performance of ethnic groups was examined. The findings for the 13 Trait scales and their relationship with ability test scores revealed some interesting findings, some of which were predicted in one form or another based on previous theoretical findings, but many of which came as something of a surprise. One finding of particular note was the apparently orthogonal nature in the dataset of the Trait scales that were conceptually related to Extraversion (Trait Sociability) and the anxiety dimension of Emotional Stability (Trait Stability), and test performance. This is surprising, given that these traits have previously demonstrated moderately strong correlations with test performance (e.g. Chamorro-Premuzic & Furnham, 2004). Taking these findings individually, the lack of association

between Stability and test performance may well be attributable to the nature of the testing environment. As the tests were administered in a low-stakes environment, far removed from that of a selection process, an individual candidate's ability to manage stressors was likely less important than it might have been had the stakes been higher. The Sociability findings, however, are more difficult to explain. Previous research would suggest that this trait should be positively correlated with test performance, greater assertiveness being associated with greater confidence in one's answers and lower levels of arousal conferring less sensitivity to external distractors (Chamorro-Premuzic & Furnham, 2004). However, Sociability displayed at best a weak *negative* correlation with test performance. One possible explanation for the lack of positive correlation between Sociability and test performance is, again, associated with the low-stakes nature of the testing session in the present study. In the pressurised environment of employment testing, stressors are likely to have a detrimental effect on attention, as has been observed by Lupien et al. (2007), through the effect of elevated glucocorticoid levels. The lower levels of arousal associated with higher Sociability scores would allow high scorers to manage this detrimental effect on attention more effectively than higher scorers. However, in the absence of these stressors, an individual's arousal levels would be largely irrelevant. This might, therefore, be able to account for a lack of positive correlation between Sociability and test performance in the present study.

By comparison, the Intellect Trait scale behaved, for the most part, as predicted. It displayed weak-moderate positive correlations with all of the ability measures in the study. This is consistent with the growing body of literature on the interface between personality and intelligence, in that Intellect is a facet of Openness, and traits related to Openness have consistently been demonstrated to predict performance on ability measures (e.g. Ackerman & Heggestad, 1997; Moutafi, Furnham & Paltiel, 2005; Woods, Bellman-Jeffries & Hinton, 2013). Current thinking in this area posits that the mechanism by which Openness influences ability is through a reciprocal relationship. Woods et al.'s (2013) Dynamic Development Model (DDM) argues that, in early childhood, Openness predicts broad

preferences to engage in a range of cognitive activities. In turn, ability predicts the likelihood of success in these activities, which influences the likelihood that a person seeks out more of these activities. This model has two important implications for the present research. Firstly, Intellect would be most strongly associated with measures of ability that were most strongly loaded on g_c, as high scorers on this scale would be more likely to display a preference for intellectual pursuits. This appears to be the case in the present study, as Intellect's strongest correlation by some margin was with the PfS Verbal (r = .32; p < .01). Secondly, according to the logic of the DDM, Intellect might also influence test performance through the additional mediating pathway of test familiarity, in that higher scorers would be likely to gain increased exposure to tests through a preference for completing them, based on their success doing so. Some support for this hypothesis was found in the present study by the moderately strong correlation between Intellect and test familiarity. Furthermore, ethnicity was found to explain meaningful variance in Intellect scores, and much of this variance could be accounted for by participant social status. This would suggest that ethnic differences in socio-economic factors give rise to differences in intellect, which, in turn, can explain a pathway to differences in test familiarity.

However, these results were perhaps not as clear cut as they could have been. If it were truly the case that test familiarity acts as a partial mediator of the relationship between Intellect and test performance, we might expect a pattern of correlations between Intellect and the ability test scores that mirrored those of test familiarity closely. However, Intellect appears to be more strongly correlated with the MAT's Speed of Working scale than it is with Decision Efficiency. This is contrary to the logic of test familiarity's effect on test performance outlined in the previous section, in which test familiarity is proposed to aid performance by encouraging an efficient balance of speed and accuracy. This warrants further research to clarify this relationship, though – if Intellect can be demonstrated to influence performance according to this mediating relationship – it would imply that Intellect

should not be viewed as a performance facilitating factor *per se*, more as an antecedent to performance facilitation.

The final finding of note when the personality scales' relationships with test performance were examined was that of the Optimism scale. Optimism, theoretically, should be related to test performance, in that it taps into both Extraversion and Emotional Stability, those who score highly on the Optimism scale being more likely to be both extraverts and emotionally stable. Furthermore, the Optimism scale has some conceptual overlap with some facets of Core Self Evaluations, particularly Locus of Control. Previous research has consistently demonstrated (e.g. Maqsud & Rouhani, 1991; Meyerhoff, 2004) that those of higher socio-economic status are more likely to have an internal locus of control (i.e. that which is represented by higher Optimism scores). In the present study, Optimism was positively correlated with both measures of social status, particularly with participants' present level of social status. Furthermore, participant social status appeared to be able to account for a degree of the variance in Optimism scores accounted for by ethnicity.

This pattern of associations would set up Optimism neatly as a partial mediator of the relationship between Ethnicity, Socio-economic Status and test performance. In terms of its association with performance, the literature on locus of control has long acknowledged the positive association that an internal locus of control has on academic achievement (e.g. Findlay & Cooper, 1983), though, as previously been mentioned in section 2.5.2, it is a dangerous assumption to generalise research findings for academic performance to those of ability test performance. However, in the present study, Optimism was found not to be positively correlated with test performance as might be expected based on this line of reasoning. In fact, the opposite was true in the present study, in that Optimism scores were *negatively* related to test performance on all ability measures. Though these effects were of trivial size for many of the measures, these negative relationships were weak-moderately strong for the PfS Abstract, the PfS Verbal and the MAT's Decision Efficiency scale. Examining the beta weights in the regression analyses summarised in section 4.3.7, an

increase of one point on this scale is associated with a decrease in raw score on these scales of between 0.25 and 0.3 S.D. units. Once possible explanation for the existence of this negative relationship comes in the form of optimism's possible association with over-confidence. Furnham, Chamorro-Premuzic and McDougall (2002) observed a detrimental effect of over-confidence on students' academic performance. It is possible that this detrimental effect of over-confidence could be generalised to the case of ability test performance. High optimism could lead a test taker to have the expectation of success when completing a test. This expectation could lead them to have the expectation that their initial answer is correct, leading them to respond to test items more quickly without the careful checking of their answers to which those of lower optimism might be predisposed. Nevertheless, based on the present study's findings, this explanation is largely conjecture, and would require further investigation in the future.

Examining how optimism, test performance and socio-economic factors vary across ethnicities more closely, a somewhat strange pattern emerged. It appeared that, for the Black and Asian groups, higher participant occupational prestige and social status did not conform to the positive trend observed in the dataset as a whole, but rather seemed negatively related to one another. This potentially offers an opportunity to reconcile the otherwise anomalous relationships observed between these variables. African and South East Asian cultures tend to be more collectivistic than individualistic (Hofstede, 1980). Those from collectivist cultures place much more emphasis on family and social bonds than do those from individualist cultures, for whom personal achievement is an important source of satisfaction. It is possible, then, that participants from ethnic groups whose origins are within these cultures might possess this more collectivist world view as a function of the values that were instilled into them during development by family members. This might account for why optimism seems positively correlated with social status and occupational prestige for the White group, but, for the other ethnic groups in the study, these two variables appear not to be related. This explanation is somewhat tenuous, though these results do

suggest a potentially interesting avenue of future research. In particular, should the negative correlation between optimism and test performance be replicable in other samples, it might potentially lead to our understanding of a hitherto discounted personality trait that could have a marked impact on performance.

5.3 Academic Contribution

Together, the findings of the present research make a number of contributions to the ethnic bias literature. Firstly, the present research provides some much needed insight into the state of ethnic group differences in test performance in the UK today. Historically, the vast majority of research in this field has focused on samples from the US. Moreover, the research considers ethnic group differences in a much broader way than it has tended to be previously. The literature overwhelmingly favours the investigation of differences between White and Black test takers, often neglecting to consider differences between the White majority and other ethnic groups. The effect of this is to somewhat diminish the importance of many minority ethnic groups for whom fair treatment in selection is no less pertinent. For example, the most recent census (ONS, 2012) revealed that, together, Indians, Pakistanis and Bangladeshis represent approximately 5.8% of the total population of England and Wales, whereas all Black ethnic subgroups together only represent 3.3%. Up until now, both academics and policy makers might have chosen to focus on ensuring that Black test takers are treated fairly in selection, given that the largest differences in test performance tend to have been observed between White and Black test takers. However, the present research indicates that – for at least some measures – White-Asian differences can be of a comparable magnitude, particularly when the performance of Asian ethnic subgroups are considered separately to one another. The present research, therefore, brings to the foreground the plight of many minority ethnic groups who have classically been discounted in the literature, but for whom the possibility of being disadvantaged during selection is a very real possibility.

A second contribution that the present research makes to the field of ethnic bias research is to offer clarification of understanding and recommendations for the use of MLVM for the investigation of DTF in ability testing data. In particular, Study 1 highlighted the impact that misspecification of a mixture model can have on the results of MLVM, as well as offering a simple approach to identifying whether latent classes identified are likely to be spurious due to misspecification. Using EFA based on tetrachoric correlations, the number of latent factors underlying the items that make up an ability test can be established. If the test is found to have a multifactorial structure (i.e. it violates the assumptions of MLVM), a series of follow-up mixture models can be run, based on the observed pattern of item loadings on each factor. These subsamples of items within the test will not violate the assumptions of MLVM. Furthermore, the number of latent classes indicated in these analyses provides insight into how reliable the findings of the main analysis are. If no heterogeneity in response patterns is observed for these subsets of items, it implies that any heterogeneity identified in the sample as a whole is likely due to model misspecification. This represents a safeguard that can be employed when investigating response behaviour in ability testing data that should lead to more reliable conclusions on the nature of DTF within these tools.

The final contribution that the study makes to the extant literature on ethnic group test performance differences is in the proportion of these differences that can be explained by non-cognitive factors. Based on the contribution of test familiarity and optimism, a very large degree of ethnic group test score differences can be accounted for. Given that differences in test familiarity can be explained – at least in part – by differences in socio-economic factors between these groups, this allows greater insight into the nature of ethnic group test performance differences than has previously been possible. This leads on to the largest contribution that the present study makes to the literature, namely the theoretical model that is proposed to provide better understanding of how test performance differences between ethnic groups come about.

5.3.1 A Model of Ethnic Group Difference in Test Performance

Based on the findings of the present research, a new model of how ethnic differences in test performance arise is proposed. This model differentiates itself from previous explanations of ethnic group test score differences in that it considers the root causes of test performance differences in low-stakes contexts. Much of the literature has focused, thus far, on factors such as test anxiety and stereotype threat (e.g. Steele & Aronson, 1995; Sackett, Hardison & Cullen, 2004; Helms, 2005; Nguyen & Ryan, 2008) that can debilitate the performance of test takers from minority ethnic groups. These effects are well understood, but offer little explanation as to why the performance of those from some ethnic groups is facilitated in testing sessions in which anxiety should not hinder performance. The model in Figure 27 below presents a synthesis of the results of the present research, offering a proposed causal explanation for how ethnic group test performance differences might be influenced by construct-irrelevant factors that have their root cause in socio-environmental factors.

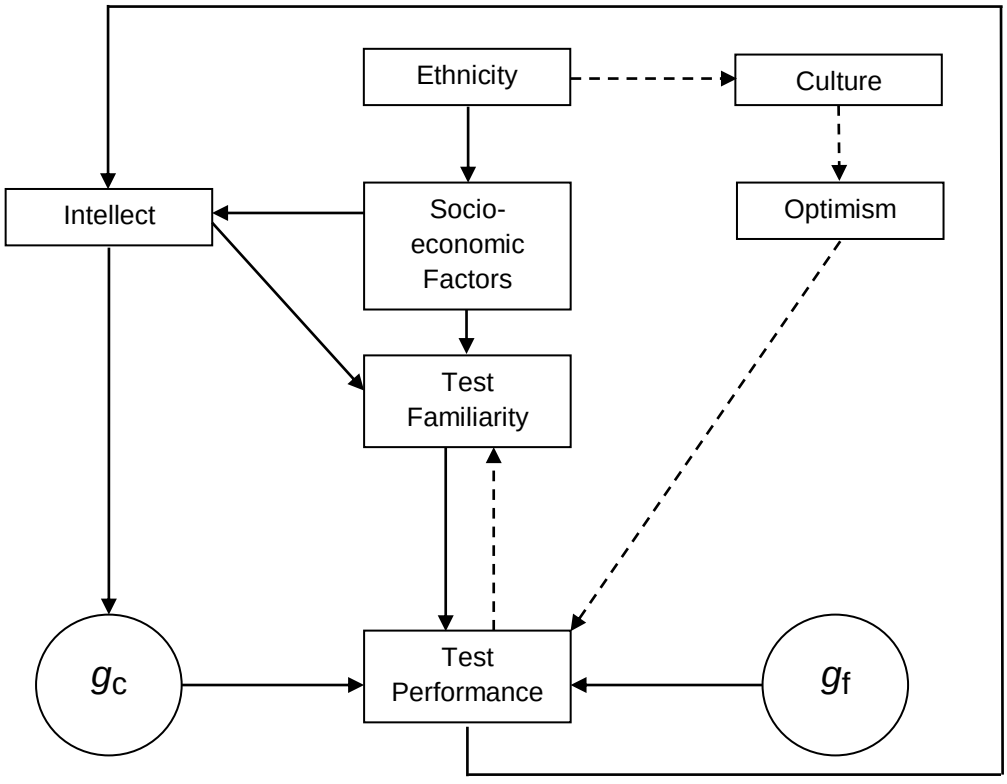

Figure 27. Proposed model showing how, in low-stakes situations, ethnicity might give rise to differences in test performance through the mediating links of socio-economic factors, culture, optimism, and test familiarity.

The model in Figure 27 shows how test performance is facilitated through a candidate's familiarity with ability testing in general. This test familiarity is strongest under two conditions: Firstly, an individual with high intellect (i.e. their preference for engaging in intellectual pursuits) is more likely to display a natural preference for activities that expose them to ability testing. Secondly, the socio-economic environment in which a person develops is likely to influence the likelihood that they come into contact with testing in their early years, high-SES individuals being more likely to encounter ability tests of some form than those from low-SES backgrounds. The performance-facilitating effect of test familiarity, therefore, has its roots in factors that depend upon socio-environmental context. Furthermore, as test familiarity increases the likelihood of success in ability tests, a positive feedback loop is created. According to the logic of the DDM, success in these early cognitive tests will encourage an individual to actively seek out more tests, further increasing their test familiarity, thus further facilitating their performance on future tests. The effect of this is that those who become familiar with cognitive tests during development are likely to increase their test familiarity as time goes on, whereas those who are not exposed to them will not. This means that those from low-SES environments – in which members of minority ethnic groups, statistically, are far more likely to have grown up – will likely have had much less of this performance-facilitating exposure than those from high-SES ones when encountering them as part of a selection process, thereby disadvantaging them.

Conversely, although it appears to be able to explain a large degree of ethnic group test performance differences, optimism does not appear to have its roots in socio-economic factors. Instead, the model proposes that – at least for ethnic groups whose cultural origin is Collectivist – these differences in optimism arise primarily from the influence of internalised familial beliefs and norms. The proposed mechanism by which optimism might debilitate test performance for those groups who have typically high levels of optimism is through the effect of over-confidence on how a candidate might approach the items within a test. However,

both of these proposed links in the model need to be verified through replication of the effects observed in the present research.

This proposed model provides a framework that can guide future research and allow better understanding of the mediating role of non-cognitive factors on the relationship between socio-environmental factors and test performance. The conceptual definition of test familiarity developed for the present study will allow future research to be more focused, ensuring that the findings of studies relating to test familiarity are not conflated with those relating to testing practice effects or academic examination familiarity. The model additionally includes a number of proposed links that could – if empirical support for them could be found – explain construct-irrelevant performance facilitation and debilitation more fully. These potential links are explored more fully in Section 5.6 below.

5.4 Practical Impact and Recommendations for Practitioners

The test familiarity scale offers a simple solution to the problem of ethnic bias in ability testing in selection. Potentially, the short, four-item scale that has been developed as part of the present research could be included as a tool to aid interpretation of ability test scores in selection processes. Then, assuming that the association between test familiarity and performance on the test is known approximately, a candidate's test score can be adjusted so that the test's assessment of their ability is closer to their true level of ability. Any remaining difference between the candidate's test score and that of others will then be more indicative of differences in true ability. In effect, this would serve to reduce the error bars around a candidate's observed score, which, in turn, would allow practitioners to be more confident when setting cut-off scores during a second sift, or deciding when test score differences between candidates are representative of differences in true ability during a late-stage selecting-in process.

Furthermore, as test familiarity can explain a large proportion of ethnic group test performance differences on a variety of cognitive measures, it provides a more desirable potential alternative to the methods for managing these differences in selection processes. Typically (as has been explored in Chapter 2), practitioners rely on two techniques for the identification of ethnic bias in selection, both of which have inherent drawbacks associated with how they then recommend that ethnic group differences are mitigated. In the case of the 4/5^{ths} Rule of Thumb, if adverse impact is identified, accepted wisdom in practice is that the test's cut score should be lowered to avoid unfairly disadvantaging members of the lower-scoring group. However, this serves to reduce the utility of the measure for cost-effective selection. In the case of differential prediction, if particular ethnic groups are shown not to share a common regression line when future job performance is regressed on to test score at selection, it is recommended that alternative, ethnicity-specific norm groups are used to score the test. Apart from introducing the possibility of litigation on the basis of direct discrimination, this presupposes that specific norm groups exist for the ethnic group in question. To complicate matters, the present research highlights the inherent problem with making adjustments to a selection method on the basis of an assumption that broad ethnic groups are homogeneous in nature. As has been demonstrated, using a single, broad ethnic norm group to represent candidates from a variety of subgroups within this ethnicity – for example, using an 'Asian' norm for ethnically Chinese, Indian and Pakistani candidates – would not be a robust (nor ethical) way of dealing with any prediction bias present in the test. The use of the test familiarity scale, then, represents a discreet transformation that can potentially be applied to scores on any ability test that is likely to increase the fairness of the use of these tools when selecting from candidate pools that are made up of at-risk ethnic groups.

It is recognised, however, that the scale needs further refinement and validation to ensure it is robust before deployment as part of high-stakes decision making. One particular issue that would need to be resolved would be how one could discourage dishonest responses to

the scale. Were it to become common knowledge that responding to the scale by portraying oneself as unfamiliar with ability testing could provide a potential advantage during selection, it would invalidate the scale as a means of reducing ethnic differences. The same issues, therefore, exist for the scale as for personality measures in terms of Socially Desirable Responding (SDR) behaviour. However, the inclusion of a measure such as an SDR scale within the tool might serve to make it unnecessarily unwieldy.

Apart from the practical application of the test familiarity scale itself, the present research also has implications at the societal level. It has been demonstrated that a large degree of the performance differences between ethnic groups depends on that group's familiarity with testing, and that the process by which test familiarity is developed begins during the early years. Therefore, if, as a society, we want to minimise ethnic group performance differences on selection tests, the key would appear to lie in ensuring that all groups of people – irrelevant of socio-economic background – are made familiar with testing from an early age, and throughout their lives. The inclusion of an introduction to the basic forms of ability testing as part of the National Curriculum would ensure that this process starts during cognitive development. This could then be built upon by the integration of ability testing into career development programmes, job seeker support programmes, and so on. The more that testing can be made part of society at all levels, the more likely it will be that ability tests can be used without disadvantaging minority social groups, whilst still continuing to be the best single predictor of future job performance available.

5.5 Limitations to the Present Research

The implications of the present research must be tempered by consideration of its shortcomings. Three of the more important of these limitations will be discussed in this section, before how these limitations might impact on the future directions for research is discussed on the following section.

A key limitation of the second study's design was in the large number of ability tests that were administered to participants during the testing session. Participants were asked to complete five ability measures, one after another, with only a relatively short comfort break halfway through the session. In most cases, the entire session took around 2 hours to administer. These five tests were presented in a consistent order for all participants. The reason for this was to minimise order effects, though this made it highly likely that participants would have suffered a degree of fatigue that might have influenced their performance on later tests such as the MAT and Raven's SPM. It has been observed in a variety of contexts that attention and task performance are adversely affected by mental fatigue during prolonged exposure to complex cognitive tasks (e.g. Boksem, Meijman & Lorist, 2005). As such, the longer participants spend completing ability tests, the less their performance will depend on the construct of interest, and the more it will depend on unrelated factors such as attention.

To illustrate the effect that this might have had on participants' performance, one may compare the group mean performance of groups on the Raven's SPM across the two studies. In the first study, most ethnic groups scored between 21 and 23 raw score points. However, when participants in the second study completed the SPM, all ethnic groups – regardless of the classification of ethnicity used – scored either slightly or substantially more poorly than this range. Furthermore, the differences in performance on the SPM between ethnic groups and the White majority seem to be exaggerated in the second study, displaying much larger values of Cohen's *d* than for equivalent groups in the first study.

This observation leads the author to recommend that practitioners exercise caution when combining multiple ability tests as part of an assessment process. If too many measures are administered to candidates at once, there is a risk that candidates from all groups may not be able to perform to the best of their ability on the later tests. The real danger, however, might present itself in the form of the increased potential for some of these later measures to display adverse impact towards minority ethnic groups. Selection ratios on the basis of

performance on these later tests, therefore, would be particularly important to monitor, given that they may not have displayed substantial ethnic group differences when they were originally validated.

The second limitation of the present research was in the lack of inclusion of a criterion of job performance. A key issue surrounding the use of ability tests in organisational contexts is the accuracy by which they can predict future job performance equivalently for different ethnic groups at the point of selection. It would have been, therefore, meaningful to interpret ethnic group test performance differences in light of any potential differences between these groups in their job performance. There were two main reasons why this was impossible, however. Firstly, the recruitment strategy used in the second study targeted participants who were UK residents of working age, though not necessarily employed. The reason for this was to avoid the problem of range restriction that is so prevalent in studies of this kind by considering the test performance of those who were unemployed (and without work) alongside that of those who were employed (gainfully or otherwise). This necessarily meant that job performance data could not be collected for all participants. Had this been possible, though, a further issue presents itself in how this job performance could be measured. The conceptualisations of bias in the literature that focus on the prediction of future performance (e.g. those of Cleary and Thorndike) rely on job performance being represented by a perfectly reliable criterion (Chung-Yan & Cronshaw, 2002). However, in practice, there is no measure of job performance that can be considered completely reliable and free from irrelevant bias.

The final limitation to be discussed lies within the study's cross-sectional nature. Correlational studies of this kind suffer – by design – from an inability to establish causality within relationships, beyond their likely nature on a conceptual level. The present study has made the claim that test familiarity increases one's performance on ability tests, most likely through an increased probability of employing an effective balance of speed and accuracy. However, due to the study's correlational nature, it is impossible to know for sure whether

this is the case, or whether, in fact, the opposite is true, that high performance on these tests caused participants to self-assess their own test familiarity more highly than they might otherwise have done. This second explanation might tie in somewhat with Chamorro-Premuzic and Furnham's (2004) concept of subjectively assessed intelligence (SAI), a construct which conceptually overlaps with test familiarity to a degree. Chamorro-Premuzic and Furnham observed SAI to be moderately positively correlated with IQ test performance, showing a similar effect size to that observed for test familiarity in the present study. They reasoned that a low SAI could potentially influence a candidate of high ability to perform less well than they were able, but also that performance had a reciprocal effect on SAI, in that candidates revise their SAI based on their previous test performance. Poor or inaccurate feedback from a test in which a candidate performed well might, therefore, influence them to revise their SAI down, impacting on their future performance. In order for a clearer differentiation to be made between test familiarity and SAI, therefore, further research would need to be conducted that examined both constructs alongside one another.

5.6 Directions for Future Research

On the basis of the findings of the present research, the author calls for a line of research to be established to further clarify the nature of performance facilitative and debilitating factors to test performance. In particular, greater understanding is needed around these factors in terms of how they arise, and how they may be either fostered or overcome to aid the performance of test takers from the broad spectrum of social backgrounds. The first potential direction for future research to be considered is a further clarification of the nature of the relationship between Chamorro-Premuzic and Furnham's (2004) concept of SAI (a construct that is heavily influenced by personality) and test familiarity as it is captured in the present study (one which is largely orthogonal to personality, with the exception of Intellect). Furnham, Moutafi and Chamorro-Premuzic (2005) have considered the relationship between similar constructs before, but only used a dichotomous, single-item measure of 'previous IQ

experience' (to which participants answered either 'yes' or 'no' to whether or not they had previously had their IQ tested). This item would not measure a participant's test familiarity *per se* according to the conceptual definition laid out in the present research. Nor would it be able to differentiate particularly well between different degrees of test familiarity. To solidify the findings of the present study, future research should investigate these two constructs side by side, firstly to investigate whether a convincing degree of discriminant validity between the two measures can be established, and – if it can – in what ways these two constructs interact with one another to influence test performance.

A second important follow-up study to this one would seek to overcome the limitations of its correlational nature. An intervention study should be conducted to examine whether test familiarity can be trained to reduce differences in test performance between ethnic groups. The effect of familial social status implies that developmental environment plays a key role in the development of test familiarity, but the scale itself does not necessarily tap into a longitudinally stable construct.

Intervention studies designed to increase test performance have been tried periodically in the past (e.g. Ryer, Schmidt & Schmidt, 1999, cited by Sacket et al., 2001) without great success. However, findings in this area must be viewed as unreliable, due to flaws in these studies' experimental designs such as the failure to include control groups. Furthermore, without a strong theoretical model of how test familiarity comes about and the way in which it influences test performance such as the one presented in the present study, these previous studies would not have been able to design their test coaching programmes as effectively as is possible on the basis of the present research. The proposed intervention study would, therefore, recruit four groups of participants. These groups would consist of two experimental groups (one White and one of a single non-White ethnicity) and two control groups (having similar ethnic make-ups to the experimental groups). All participants would be tested initially on a short measure of ability and their baseline test familiarity would be recorded. The control groups would then be given some unrelated training, while the

experimental groups would be given a short, intensive programme of test familiarity training, made up of exposure to tests, instruction in the general forms of test questions and their cognitive rules, and – most importantly – test skills training focusing on optimising each participant's balance of speed and accuracy when completing these tests. After this, all participants would again assess their test familiarity (to ascertain whether the relevant training had led to gains on this construct) and would complete another ability test (to assess whether any gain in test familiarity in the short term translated into measureable test performance gains). The implication of finding that test familiarity can be trained could allow for the wide-scale implementation of familiarity training, providing a simple, cost effective way of addressing at least some of the persistent ethnic group differences observed in ability testing, and allowing their fairer deployment in organisational settings. Furthermore, this would allow the causality of the relationship between test familiarity and test performance to be properly established.

Beyond the influence of test familiarity on test performance, future research should additionally explore some of the other performance facilitating and debilitating factors considered in the present study more fully. In particular, the relationship of optimism to test performance would need to be replicated, and its link to internalised cultural norms established. If this were possible, it would be interesting to explore how this trait interacted with related facets of Core Self Evaluations such as locus of control and self-efficacy in the facilitation of test performance. Additional work would also benefit from the consideration of the role of test anxiety on performance debilitation, not only by measuring trait emotional stability, but also by assessing how this translates into actual anxiety as a state. To do this, testing would need to be conducted in a more high-stakes situation, one in which test anxiety is more likely to influence performance than in the present study. Assessing test anxiety using a measure such as the State-Trait Anxiety Inventory (STAI; Spielberger, Gorsuch & Lushene, 1983) would allow the relative impact of these performance debilitating factors to be assessed alongside the performance facilitating effect of test familiarity when attempting

to explain ethnic group test performance differences under circumstances that more closely matched those under which candidates would likely encounter ability testing in organisations.

5.7 Conclusion

The present research identified two key knowledge gaps in the ethnic bias literature. The first of these was the apparent inability for DTF to account for the observed ethnic group test performance differences that have been so frequently observed in the literature. Using up-to-date data and robust identification techniques, ethnic DTF was unable to be demonstrated in either of the Raven's measures, even though substantial ethnic group test score differences were observed. From this, one can conclude that investigating ethnic DTF using IRT methodology cannot substantially ameliorate the current issues facing selection assessment in terms of ensuring that ability tests do not disadvantage minority social groups.

Conversely, the present research made substantial progress in identifying two key non-cognitive factors that, alone, can explain a very large proportion of ethnic differences on a variety of ability. The findings of the research make a reasoned argument not only for their existence and effect on test performance, but also of their root causes and the mechanism by which they facilitate performance. In so doing, recommendations can be made for selection practitioners about how ethnic differences should be viewed in ability test performance, and how, potentially, it could be mitigated. A number of potentially important avenues for future research have been identified. The author closes with an appeal to the field to collaborate in moving forward with new research in to the nature of the phenomena, that, one day, we might be able to use ability testing as a robust selection method without the ever-present worry of unfairly excluding those we have, as a society, sought to include.

Overall, the present research represents a significant step towards the resolution of problems that have dogged the field of personnel selection and assessment for as long as it has existed. While it is – and should only be viewed as – a preliminary step towards understanding a key differentiator of test performance between ethnic groups, it represents hope for the future. The threat of unfairness and bias in selection testing has, previously, cast serious doubts on whether assessments of this kind should be used in modern, multicultural societies. As the UK continues to become more ethnically diverse, the need to assess all groups of people without any of them being disadvantaged is becoming as important an issue as the accurate prediction of future performance in selection. As our understanding of how these differences come about increases, so does our ability to account for them. The research that the present study influences will provide insights that will, one day, allow organisations to accurately predict performance without creating barriers to employment for any social group.

LIST OF REFERENCES

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence*, 22 (2), 227–257.

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin*, 121, 219–245.

Ackerman, P. L., Kanfer, R., & Goff, M. (1995). Cognitive and non-cognitive determinants and consequences of complex skill acquisition. *Journal of Experimental Psychology*, 1, 270–304.

Adams, J., & Weakliem, D. L. (2011). August B. Hollingshead's "Four Factor Index of Social Status": From Unpublished Paper to Citation Classic. *Yale Journal of Sociology*, 8, 11–19.

Aguinis, H., Culpepper, S. A., & Pierce, C.A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology*, 95 (4), 648–680.

Anastasi, A. (1981). Coaching, test sophistication and developed abilities. *American Psychologist*, 36, 1086−1093.

Anderson, R. C., & Pearson, P. D. (1988). A schema-theoretic view of basic processes in reading comprehension. In P. L. Carrell, J. Devine, & D. E. Eskey (Eds.). *Interactive Approaches to Second Language Reading*. Cambridge University Press.

American Psychological Association, Task Force on Socioeconomic Status. (2007). *Report of the APA Task Force on Socioeconomic Status*. Washington, DC: American Psychological Association.

Arnold, J., Silvester, J., Patterson, F., Robertson, I., Cooper, C., & Burnes, B. (2005) *Work Psychology* (4th Edition). Harlow: Pearson Education Ltd.

Ashton, M. C., & Lee, K. (2006). "Minimally biased" *g*-loadings of crystallized and non-crystallized abilities. Intelligence, 34 (5), 469−477.

Aston Business Assessments Ltd (2011). *Trait: Manual and User Guide*. Birmingham, UK.

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling: A Multidisciplinary Journal*, 21 (1), 102−116.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.

Barro, R. J., & Lee, J. W. (2001). International data on educational attainment: updates and implications. *Oxford Economic Papers*, 53 (3), 541–563.

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model*. Princeton, NJ: Educational Testing Service.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: potential problems and promising opportunities. *Psychological Methods*, 9 (1), 3–29.

Beauducel, A., Brocke, B., & Liepmann, D. (2001). Perspectives on fluid and crystallized intelligence: facets for verbal, numerical, and figural intelligence. *Personality and individual differences*, 30 (6), 977–994.

Berry, C. M., Clark, M. A., & McClure, T. K. (2011). Racial/ethnic differences in the criterion-related validity of cognitive ability tests: A qualitative and quantitative review. *Journal of Applied Psychology*, 96 (5), 881–906.

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology*, 78 (3), 387–409.

Bobko, P. & Roth, P. L. (2009) An Analysis of Two Methods for Assessing and Indexing Adverse Impact: A Disconnect Between the Academic Literature and Some Practice. In J.L. Outtz (Ed) *Adverse Impact: Implications for Organizational Staffing and High Stakes Selection*. New York: Routledge.

Boksem, M. A., Meijman, T. F., & Lorist, M. M. (2005). Effects of mental fatigue on attention: an ERP study. *Cognitive Brain Research*, 25 (1), 107–116.

Borsboom, D. (2006a). When does measurement invariance matter? *Medical Care*, 44 (11), 176–181.

Borsboom, D. (2006b). The attack of the psychometricians. *Psychometrika*, 71 (3), 425–440.

Bouchard, T. J., & McGue, M. (2003). Genetic and environmental influences on human psychological differences. *Journal of Neurobiology*, 54 (1), 4–45.

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist*, 39 (3), 214–227.

British Psychological Society (2010). *Code of Human Research Ethics*. Leicester: BPS.

British Psychological Society (2011). *Test User's Handbook: Information on the British Psychological Society's qualifications in psychological testing (v0.3)*. Leicester: BPS.

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, 27(4), 509–540.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: a theoretical account of the processing in the Raven Progressive Matrices Test. Psychological review, 97 (3), 404–431.

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-analytic Studies*. Cambridge University Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, 54 (1), 1–22.

Chamorro-Premuzic, T., & Furnham, A. (2004). A possible model for understanding the personality-intelligence interface. *British Journal of Psychology*, 95 (2), 249–264.

Childs, R., Gosling, J., Parkinson, M., & McDonald, A. S. (2013). *Profiling for Success: Reasoning Tests User's Guide v1.3*. Maidenhead, Berkshire: Team Focus.

Chung-Yan, G. A., & Cronshaw, S. F. (2002). A critical re-examination and analysis of cognitive ability tests using the Thorndike model of fairness. *Journal of Occupational and Organizational Psychology*, 75 (4), 489–509.

Chartered Institute of Personnel and Development (2013). *Annual Survey Report 2013: Resourcing and Talent Planning*. Retrieved on 26/9/14 from http://www.cipd.co.uk/binaries/resourcing-and-talent-planning_2013.PDF.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, 5 (2), 115–124.

Cook, M. (1998). *Personnel Selection: Adding Value Through People* (3rd Edition). Wiley & Sons.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Routledge Academic.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, 42 (2), 133–148.

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, 30 (2), 163–183.

Cooper, D., & Robertson, I. (1995). *The Psychology of Personnel Selection: A Quality Approach*. London: Routledge.

Cronbach, L. J. (1990). *Essentials of Psychological Testing* (5th Edition). New York: Harper Collins.

Davis, J., Smith, T., Hodge, R., Nakao, K., & Treas, J. (1991). *Occupational Prestige Ratings from the 1989 General Social Survey*. Ann Arbor MI: Inter-university Consortium for Political and Social Research.

Dean, M. A., Roth, P. L., & Bobko, P. (2008). Ethnic and gender subgroup differences in assessment center ratings: a meta-analysis. *Journal of Applied Psychology*, 93 (3), 685–691.

DerSimonian, R., & Laird, N. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review*, 18, 694–734.

DeShon, R. P., Chan, D., & Weissbein, D. A. (1995). Verbal overshadowing effects on Raven's Advanced Progressive Matrices: Evidence for multidimensional performance determinants. *Intelligence*, 21, 135–155.

Dickens, W. T. & Flynn, J. R. (2006). Black Americans Reduce the Racial IQ Gap: Evidence From Standardization Sample. *Psychological Science*, 17 (10), 913–920.

DiStefano, C., & Morgan, G. B. (2014). A Comparison of Diagonal Weighted Least Squares Robust Estimation Techniques for Ordinal Data. *Structural Equation Modeling: A Multidisciplinary Journal*, 21 (3), 425–438.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89 (426), 463–475.

Ek, E., Remes, J., & Sovio, U. (2004). Social and developmental predictors of optimism from infancy to early adulthood. *Social Indicators Research*, 69 (2), 219–242.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology*, 128, 309–331.

Equal Employment Opportunity Commission (n.d.). *Uniform Employee Selection Guidelines: Interpretation and Clarification (Questions and Answers).* Retrieved on 16/3/2014 from http://uniformguidelines.com/qandaprint.html#2.

ESRC (2012). *ESRC Framework for Research Ethics (FRE) 2010: Updated September 2012.* Retrieved on 21/8/2014 from http://www.esrc.ac.uk/_images/framework-for-research-ethics-09-12_tcm8-4586.pdf.

Fidalgo, A. M., Mellenbergh, G. J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online*, 5 (3), 1–11.

Findley, M. J., & Cooper, H. M. (1983). Locus of control and academic achievement: A literature review. *Journal of Personality and Social Psychology*, 44(2), 419–427.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with Logistic Regression for Differential Item Functioning Detection. *Educational and Psychological Measurement*, 67, 373–393.

Frierson, H.T. (1986). Enhancing minority college students' performance on educational tests. *Journal of Negro Education*, 55, 38–45.

Furnham, A., Chamorro-Premuzic, T., & McDougall, F. (2002). Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance. *Learning and Individual Differences*, 14 (1), 47–64.

Furnham, A., Moutafi, J., & Chamorro-Premuzic, T. (2005). Personality and intelligence: Gender, the Big Five, self-estimated and psychometric intelligence. *International Journal of Selection and Assessment*, 13 (1), 11–24.

Gibson, S. G., & Harvey, R. J. (2003). Gender and ethnicity based differential item functioning on the Armed Services Vocational Aptitude Battery. *Equal Opportunities International*, 22 (4), 1–15.

Goldberg, L. R., Sweeney, D., Merenda, P. F., & Hughes Jr, J. E. (1998). Demographic variables and personality: The effects of gender, age, education, and ethnic/racial status on self-descriptions of personality attributes. *Personality and Individual Differences*, 24 (3), 393–403.

Gómez-Benito, J., Hidalgo, M.D., & Guilera, G. (2010). Bias in measurement instruments. Fair tests. *Papeles del Psicólogo*, 31 (1), 75–84

Gottfredson, L. S. (2005). What if the hereditarian hypothesis is true? *Psychology, Public Policy, and Law*, 11, 311–319.

Haddock, C. K., Rindskopf, D., & Shadish, W. R. (1998). Using odds ratios as effect sizes for meta-analysis of dichotomous data: A primer on methods and issues. *Psychological Methods*, 3, 339–353.

Hair, J. F., Tatham, R. L., Anderson, R. E., & Black, W. (1998). *Multivariate Data Analysis* (5th Edition). London: Prentice-Hall.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: a meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, 92 (2), 373–385.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, 76 (4), 408–420.

Heinonen, K., Räikkönen, K., Matthews, K. A., Scheier, M. F., Raitakari, O. T., Pulkki, L., & Keltikangas-Järvinen, L. (2006). Socioeconomic Status in Childhood and Adulthood: Associations With Dispositional Optimism and Pessimism Over a 21-Year Follow-Up. *Journal of Personality*, 74 (4), 1111–1126.

Helms, J. E. (2005). Stereotype threat might explain the Black-White test-score difference. *American Psychologist*, 60 (3), 269–270.

Herrnstein, R. J., & Murray, C. (1994). *The Bell Curve.* London: Simon and Schuster.

Higuera, L. A.-Z. (2001). Adverse impact in personnel selection: The legal framework and test bias. *European Psychologist*, 6 (2), 103–111.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, 21 (5), 967–988.

Hofstede, G. H. (1980). *Culture's Consequences*. Beverly Hills, California: Sage.

Hollingshead, A. B. (1975). Four Factor Index of Social Status. Unpublished working paper, Department of Sociology, Yale University, New Haven, CT.

Horn, J. L. (1976). Human abilities: A review of research and theory in the early 1970's. *Annual Review of Psychology*, 27, 437–486.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment*, 9(1-2), 152–194.

Hunter, J. E., & Schmidt, F. L. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law*, 6 (1), 151–158.

Iacobucci, D., Saldanha, N., & Deng, X. (2007). A meditation on mediation: Evidence that structural equations models perform better than regressions. *Journal of Consumer Psychology*, 17(2), 139–153.

Ibarra, R. A. (2001). *Beyond Affirmative Action: Reframing the Context of Higher Education*. University of Wisconsin Press.

Idaszak, J. R., & Drasgow, F. (1987). A revision of the Job Diagnostic Survey: Elimination of a measurement artifact. *Journal of Applied Psychology*, 72, 69–74.

International Labour Organisation (2004). Labour Statistics Volume 3: Economically active population, employment, unemployment and hours of work (household surveys). Retrieved on 20/8/2014 from http://laborsta.ilo.org/applv8/data/SSM3/E/SSM3.html.

Jackson, D. N., & Rushton, J. P. (2006). Males have greater g: Sex differences in general mental ability from 100,000 17- to 18-year-olds on the Scholastic Assessment Test. *Intelligence*, 34 (5), 479–486.

Jensen, A. R. (1985) The Nature of the Black–White Difference on Various Psychometric Tests: Spearman's hypothesis. *Behavioral and Brain Sciences*, 8, 193–219.

Jensen, A. R. (1998). *The g Factor: The Science of Mental Ability*. Westport, CT: Praeger.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, 14, 329–349.

Johnson, J., Truxillo, D., Erdogan, B., Bauer, T., & Hammer, L. (2009). Perceptions of Overall Fairness: Are Effects on Job Performance Moderated by Leader-Member Exchange? *Human Performance*, 22 (5), 432–449.

Jonassaint, C. R., Siegler, I. C., Barefoot, J. C., Edwards, C. L., & Williams, R. B. (2011). Low life course socioeconomic status (SES) is associated with negative NEO PI-R personality patterns. *International Journal of Behavioral Medicine*, 18 (1), 13–21.

Jorm, A. F., Anstey, K. J., Christensen, H., Rodgers, B. (2004). Gender differences in cognitive abilities: The mediating role of health state and health habits. *Intelligence*, 32 (1), 7–23.

Juárez, F., & Contreras, F. (2012). The influence of optimism and socioeconomic characteristics on leadership practices. *International Journal of Psychological Research*, 5 (2), 18–29.

Kenny, E. J., & Briner, R. B. (2007). Ethnicity and behaviour in organizations : A review of British research. *Journal of Occupational and Organizational Psychology*, 80, 437–457.

Kraiger, K., & Ford, J. K. (1985). A meta-analysis of ratee race effects in performance ratings. *Journal of Applied Psychology*, 70 (1), 56–65.

Kuncel, N. R., Ones, D. S., & Sackett, P. R. (2010). Individual differences as predictors of work, educational, and broad life outcomes. *Personality and Individual Differences*, 49 (4), 331–336.

Lazarus, R. S. (1966). *Psychological Stress and the Coping Process*. New York: McGraw-Hill.

Le Fevre, M., Kolt, G. S., & Matheny, J. (2006). Eustress, distress and their interpretation in primary and secondary occupational stress management interventions: which way first? *Journal of Managerial Psychology*, 21 (6), 547–565.

Li, Y., Cohen, A., & Ibarra, R. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing*, 4 (2), 115–136.

Liao, W. W., Ho, R. G., Yen, Y. C., & Cheng, H. C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality: An International Journal*, 40 (10), 1679–1694.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods*, 38(1), 88–91.

Lupien, S. J., Maheu, F., Tu, M., Fiocco, A., & Schramek, T. E. (2007). The effects of stress and stress hormones on human cognition: implications for the field of brain and cognition. *Brain and Cognition*, 65 (3), 209–237.

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement*, 37 (4), 304–315.

Maqsud, M., & Rouhani, S. (1991). Relationships between socioeconomic status, locus of control, self-concept, and academic achievement of Batswana adolescents. *Journal of Youth and Adolescence*, 20 (1), 107–114.

Martocchio, J. J., & Whitener, E. M. (1992). Fairness in personnel selection: A meta-analysis and policy implications. *Human Relations*, 45 (5), 489–506.

Mason, D. (2000). *Race and Ethnicity in Modern Britain* (2nd Edition). Oxford University Press.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, 58, 593–614.

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology*, 38 (1), 115–142.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. (2007). Situational judgment tests, response instructions, and validity: a meta-analysis. *Personnel Psychology*, 60 (1), 63–91.

McLachlan, G., & Peel, D. (2004). *Finite Mixture Models*. New York: Wiley.

Meyerhoff, M. (2004). Locus of control: Perspective on parenting. *Pediatrics of Parents*, 21 (10), 2–8.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, 2 (3), 248–260.

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, 72 (4), 461–473.

Millward, L. (2005) *Understanding occupational and organizational psychology*. London: Sage.

Moutafi, J., Furnham, A., & Paltiel, L. (2005). Can personality factors predict intelligence? *Personality and Individual Differences*, 38 (5), 1021–1033.

Muthén, L.K. & Muthén, B.O. (2010). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

Nguyen, H. H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, 93 (6), 1314–1334.

Noble, K. G., Norman, M. F., & Farah, M. J. (2005). Neurocognitive correlates of socioeconomic status in kindergarten children. *Developmental Science*, 8 (1), 74–87.

Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: big five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology*, 93 (1), 116–130.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3 (1), 1–18.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14 (4), 535–569.

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instrumentation, and Computers*, 32, 396–402.

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences*, 43, 971–990.

Office for National Statistics (2012). *Ethnicity and National Identity in England and Wales 2011.* Retrieved on 17/7/2014 from http://www.ons.gov.uk/ons/dcp171776_290558.pdf.

Office for National Statistics (2013). *Harmonised Concepts and Questions for Social Data Sources. Primary Standards: Ethnic Group*. Retrieved on 3/8/2014 from http://www.ons.gov.uk/ons/guide-method/harmonisation/primary-set-of-harmonised-concepts-and-questions/ethnic-group.pdf.

Oliveri, M. E., Olson, B. F., Ercikan, K., & Zumbo, B. D. (2012). Methodologies for investigating item-and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing*, 12 (3), 203–223.

Ones, D. S., & Anderson, N. (2002). Gender and ethnic group differences on personality scales in selection: Some British data. *Journal of Occupational and Organizational Psychology*, 75 (3), 255–276.

Outtz, J. L. (2002) The role of cognitive ability tests in employment selection. *Human Performance*, 15, 161–171.

Owen, K. (1992). The suitability of Raven's Standard Progressive Matrices for various groups in South Africa. *Personality and Individual Differences*, 13 (2), 149–159.

Parry, C. D., & McArdle, J. J. (1991). An applied comparison of methods for least-squares factor analysis of dichotomous variables. *Applied Psychological Measurement*, 15 (1), 35–46.

Phinney, J. S. (1990). Ethnic identity in adolescents and adults: review of research. *Psychological Bulletin*, 108 (3), 499–514.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, 36, 717–731.

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, 25, 111–163.

Ree, M. J., & Carretta, T. R. (1998). General cognitive ability and occupational performance. In C. L. Cooper, & I. T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (Vol. 13, 159–184). Chichester, England: Wiley

Ree, M. J., & Earles, J. A. (1991). Predicting training success: not much more than g. *Personnel Psychology*, 44, 321–332.

Reeve, C. L., & Bonaccio, S. (2009). Measurement reliability, the Spearman-Jensen Effect and the revised Thorndike model of test bias. *International Journal of Selection and Assessment*, 17 (1), 61–68.

Reeve, C. L., Heggestad, E. D., & Lievens, F. (2009). Modeling the impact of test anxiety and test familiarity on the criterion-related validity of cognitive ability tests. *Intelligence*, 37 (1), 34–41.

Rietveld, C. A., Esko, T., Davies, G., Pers, T. H., Turley, P., Benyamin, B., ... & Ward, M. E. (2014). Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proceedings of the National Academy of Sciences*, 111 (38), 13790–13794.

Rindermann, H. (2007). The g-factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ-tests across nations. *European Journal of Personality*, 21 (5), 667–706.

Robertson, I., & Smith, M. (2001) Personnel selection. *Journal of Occupational and Organizational Psychology*, 74, 441–472.

Robertson, I., Bartram, D., & Callinan, M. (2002). Personnel Selection & Assessment. In Warr, P.B. (Ed) *Psychology at Work* (5th Edition). Harmondsworth: Penguin.

Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence*, 35 (1), 83–92.

Roth, P. L., Bobko, P., & Switzer, F. S. (2006). Modelling the behavior of the 4/5ths rule for determining adverse impact: reasons for caution. *Journal of Applied Psychology*, 91 (3), 507–522.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, 11, 235–294.

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American–White differences on cognitive tests. *American Psychologist*, 59, 7–13.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist*, 56 (4), 302–318.

Salkind, N. J. (2010). *Encyclopedia of Research Design*. London: Sage.

Samuelsen, K. M. (2008). Examining differential item functioning from a latent mixture perspective. In G. R. Hancock, & K. M. Samuelsen (Eds.) *Advances in Latent Variable Mixture Models*. Charlotte, NC: IAP.

Sawatzky, R., Ratner, P. A., Kopec, J. A., & Zumbo, B. D. (2012). Latent variable mixture models: a promising approach for the validation of patient reported outcomes. *Quality of Life Research*, 21 (4), 637–650.

Schmitt, N., Clause, C. S., & Pulakos, E. D. (1996). Subgroup differences associated with different measures of some common job-relevant constructs. *International Review of Industrial and Organizational Psychology*, 11, 115–140.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124 (2), 262–274.

Schmidt, F. L., & Hunter, J. E. (2004) General Mental Ability in the World of Work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, 86, 162–173.

Schutte, J. W., Valerio, J. K., & Carrillo, V. (1996). Optimism and socioeconomic status: A cross-cultural study. *Social Behavior and Personality: An International Journal*, 24 (1), 9–18.

Schweizer, K., Goldhammer, F., Rauch, W., & Moosbrugger, H. (2007). On the validity of Raven's matrices test: Does spatial ability contribute to performance? *Personality and Individual Differences*, 43 (8), 1998–2010.

Selye, H. (1956). *The Stress of Life*. New York: McGraw-Hill.

Sewell, W. H., & Shah, V. P. (1967). Socioeconomic status, intelligence, and the attainment of higher education. *Sociology of Education*, 40 (1), 1–23.

Smith, M. G. (1986). Pluralism, race and ethnicity in selected African countries. In J. Rex, & D. Mason (Eds.) *Theories of Race and Ethnic Relations*. Cambridge: Cambridge University Press.

Spearman, C. (1927). *The Abilities of Man*. Oxford: Macmillan.

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology*, 89 (3), 497–508.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69 (5), 797–811.

Talent Lens (2009) *Raven's Progressive Matrices*. Retrieved on 13/6/2012 from http://www.talentlens.co.uk/select/ravens-apm-and-spm-short-forms.

Team Focus (2011). *The Memory and Attention Test Manual*. Maidenhead, Berkshire: Team Focus.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods*, 16(2), 209–220.

Tran, U. S., & Formann, A. K. (2009). Performance of parallel analysis in retrieving unidimensionality in the presence of binary data. *Educational and Psychological Measurement*, 69, 50–61.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, 28, 127–154.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, 37 (3), 498–505.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. Journal of Applied Psychology, 81, 557–574.

Warr, P. B. (2001). Age and work behaviour: Physical attributes, cognitive abilities, knowledge, personality traits, and motives. *International Review of Industrial and Organizational Psychology*, 16, 1–36.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance*, 21 (3), 291–309.

Woods, S. A., Bellman-Jeffries, J., & Hinton, D. P. (2013) The Relationship between g and the Latent Structure of the CPI. Poster session presented at the BPS Division of Occupational Psychology Annual Conference 2013, Chester, UK.

Woods, S. A., Hardy, C. &, Guillaume, Y. R. F. (working paper). Cognitive Ability Testing and Adverse Impact: Meta-analytic evidence of reductions in Black-White differences in ability test scores over time.

Woods, S. A., Lievens, F., De Fruyt, F., & Wille, B. (2013), Personality across working life: The longitudinal and reciprocal influences of personality on work. *Journal of Organizational Behavior*, 34 (1), 7–25.

Yang, C.C. (2006). Evaluating latent class analysis models in qualitative phenotype identification. *Computational Statistics & Data Analysis*, 50 (4), 1090–1104.

Yen, Y.-C., Ho, R.-G., Laio, W. W., Chen, L. J., & Kuo, C. C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. *Applied Psychological Measurement*, 36, 75–87.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Generations Journal of the American Society on Aging*, 4 (2), 223–233.

Zumbo, B.D., & Gelin, M.N. (2005). A matter of test bias in educational policy research: bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies*, 5, 1–23.

Zwick, R. (2004). *Rethinking the SAT: The future of standardized testing in university admissions*. London: RoutledgeFalmer