

The Construct Validity of the NEO PI-R Personality Inventory in High Stakes Employee Selection

By Gerry Fahey, B.E., B.A., M.B.A., M.A., M.Sc.

Dublin City University Business School

July 2017

Research Supervisors: Dr. Finian Buckley

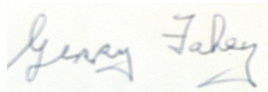
Dr. Janine Bosak

A Thesis Submitted to Dublin City University Business School
in Partial Fulfilment of the Requirements for the Degree of
Doctor of Philosophy.

Declaration

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

Signed

A handwritten signature in blue ink, appearing to read 'Genay Fakay', on a light-colored rectangular background.

ID No: 12212420

Date: 28.07.17

Acknowledgements

First of all I would like to thank my supervisor Dr. Finian Buckley for his direction and advice on this journey. Indeed there were times when I had firmly resolved to end the journey before completion but a conversation with Finian in his office always had a calming effect on me. Secondly, and just as important, without Dr. Janine Bosak's invaluable assistance in keeping me on track and focussed I would have gotten lost on the journey. The input from both was essential to get me to the finishing line. Gerry Coynygham also provided me with invaluable help in dealing with the methodological issues that arose during the journey. I owe a special word of thanks to Margaret Galuszyska who helped me deal with the admin issues as they arose. I should also mention that there are a number of academics that I had the good fortune to encounter during my years of third level education that influenced my choice of research topic and the approach that I took. To those un-named individuals I am also very grateful.

At a more personal level my three children – Paul, Julie, and Frances – each of whom in their own special way encouraged me on this path. Welcome breaks in Escorial with Paul and his family as well as the yearly family skiing holiday with my two daughters didn't go amiss. Interacting with my four grandchildren I found to be a great antidote to the tendency to the danger of becoming overly obsessive with my research topic. Helping Frances prepare for her Leaving Certificate Maths and Physics, while doing my research, was as much a help to me as it was to her. There are also three people of some decades of lasting friendship to whom I also owe much in getting to this journey's end – Bob Roeder, Eric McGrath, and Chantal Ladias.

Table of Contents

<i>Declaration</i>		<i>i</i>
<i>Acknowledgements</i>		<i>ii</i>
<i>Table of Contents</i>		<i>iii</i>
<i>List of Figures</i>		<i>vii</i>
<i>List of Tables</i>		<i>viii</i>
<i>Abstract</i>		<i>x</i>
Chapter 1	Introduction	1
Chapter 2	The Concept of Construct Validity	9
2.1	The pre 1955 Fragmented Model	10
2.2	The Unified Model	12
2.2.1	Cronbach and Meehl on Construct Validity	13
2.2.2	Loevinger on Construct Validity	16
2.2.3	Campbell and Fiske and the Multitrait-Multimethod Approach	19
2.2.4	Messick's and Embretson's Contributions	21
2.3	The Importance of Construct Validation	26
Chapter 3	The Dimensions of Personality	32
3.1	The Big Five Dimensions of Personality	33
3.2	Personality and Behaviour in the Workplace	35
3.3	The Higher Order Structure of Personality	40
3.3.1	What are the Higher Order Factors?	42
3.3.2	The Higher Order Structure is Unbalanced	46
3.4	The Lower Order Structure of the Big Five	48
3.5	The Evidence for a Hierarchical Structure	50
3.5.1	Review of the Research in support of a GFP	51
3.5.2	Relevance to the Research programme	56

3.6	Additional Measurement Issues	59
Chapter 4	Socially Desirable Responding	64
4.1	The Effect of Socially Desirable Responding on the Measurement of the Big Five	65
4.1.1	The Case against a Socially Desirable Responding Effect	68
4.1.1.1	The Case for Socially Desirable Responding in the form of Faking Good	72
4.1.2	To what extent does Faking Good occur?	76
4.2	Faking Good is a form of Moral Hypocrisy	82
4.2.1	Batson's Research on Moral Hypocrisy	84
4.2.2	Mazar's Dishonesty Research Paradigm	90
4.2.3	Behavioural Economics Research and Moral Hypocrisy	95
4.3	Remedies for Dealing with Faking Good	98
4.4	Conclusions for the Research Reviewed	102
Chapter 5	Accounting for Impression Management	109
5.1	Lie and Related Scales	110
5.1.1	Unidimensional Lie Scales	112
5.1.2	Paulhus's Socially Desirable Responding Measure	114
5.1.3	Construct Validity and the BIDR	115
5.2	Rank Order Selection Effects	123
5.2.1	Empirical Evidence for a Rank Order Effect	126
Chapter 6	Methodology Issues	131
6.1	Separating Substantive and Method Effects	133
6.1.1	Review of Extant MTMM Studies on the Higher Order Structure of Personality	138
6.1.1.1	Extant MTMM Studies	138
6.2	Factor Analytic Considerations	145
6.3	The Construct Validity of the BIDR IM Scale	152

6.3.1	Restriction of Range Issues	153
6.3.2	The Item Transparency of the Impression Management Measure	159
6.3.3	The Context Effect on the Construct Validity of the Bespoke BIDR-IM Measure used	161
6.4	Monte Carlo Simulation	167
6.5	The Research Hypotheses Tested	173
Chapter 7	Research Methods	177
7.1	Participants	178
7.2	Measures	178
7.3	Procedure	182
7.4	Analyses	183
7.4.1	Factor Analysis	184
7.4.2	IM Measure Cut Off Score	185
7.4.3	Cluster Analysis	187
7.4.4	CFA Invariance Analysis	189
7.4.5	Monte Carlo Simulations	190
7.4.6	Comparison with the Rosse, Stecher, Miller, and Levin (1998) Study	193
Chapter 8:	Results	195
8.1	Managerial Field Study	197
8.1.1	Descriptive Statistics	197
	8.1.1.1. Comparison of Managerial and Validation Samples Descriptive Statistics	201
8.1.2	Exploratory Factor Analysis	203
8.1.3	Confirmatory Factor Analysis	207
	8.1.3.1 Comments on CFA Models Tested	210
	8.1.3.2 Just Identified Model Comparison	218
	8.1.3.3 Invariance Analysis	221
8.2	Cluster Analysis	224
8.3	Monte Carlo Simulations	230
8.3.1	Descriptive Statistics	230
8.3.2	Simulation Results	232
8.3.3	Rosse, Stecher, Miller, and Levin Study comparison	236

Chapter 9:	Discussion	240
9.1	Establishing the Construct Validity of the NEO-PIR	244
9.1.1	The Higher Order Structure of the Big Five	245
9.1.2	The Link between Faking Good and the Higher Order Structure of the Big Five	253
9.2	Construct Validity of the Bespoke BIDR-IM Measure	258
9.2.1	The Practical Implications of Construct Validity	262
9.3	The Psychology of Faking Good	266
9.4	Limitations and Suggestions for future research	271
9.5	Conclusions	277
References		280

List of Figures

- | | |
|-----------|--|
| Figure 1 | Embretson's Universal System for Construct Validity |
| Figure 2 | Big Five Higher Order Putative Structure, with First Order Factors Stability and Plasticity loading on a General Factor of Personality |
| Figure 3 | Big Five Higher Order Putative Structure, with Stability and Plasticity, and no First Order General Factor of Personality |
| Figure 4 | Big Five Higher Order Putative Structure, with the Big Five Factors loading on a First Order General Factor of Personality |
| Figure 5 | Markon et al.'s (2005) Unbalanced Big Five Higher Order Structure |
| Figure 6 | Nomological Net for the Paulhus BIDR IM Scale |
| Figure 7 | Flow Chart for Results Section of Chapter 8 |
| Figure 8 | Illustration of Model 7 tested in Confirmatory Factor Analysis |
| Figure 9 | Two Higher Order Factor Model – Model Predicted Item Covariance Matrix (Hoffman, 2017) |
| Figure 10 | Plot of Frequency of Occurrence against Impression Management Scores for Cluster 1 compared to full sample of participants |
| Figure 11 | Plot of Frequency of Occurrence against Impression Management Scores for Cluster 2 compared to full sample of participants |

List of Tables

Table 1	Comparison of NEO-PIR and HPI facet loadings
Table 2	Primary and secondary factor loadings of the Facets of Extraversion and Conscientiousness
Table 3	Descriptive Statistics for the Full Sample of 443 Participants
Table 4	Comparison of Big Five Mean Scores
Table 5	Intercorrelations between the Big Five dimensions
Table 6	Intercorrelations between the Big Five Dimensions of the Managerial Sample and the Validation Sample for participants with IM scores < 12
Table 7	Means and Standard Deviations for the Big Managerial and the Validation Samples
Table 8	EFA Variance Accounted for by Factors with an eigenvalue > 1
Table 9	EFA Factor Loadings on the Two Higher Order Factors
Table 10	Variance Accounted for by Factors with an eigenvalue > 1 in the Validation and Managerial Samples for participants with an IM score <12
Table 11	EFA Factor Loadings on the Two Higher Order Factors in the Validation and Managerial Samples for participants with an IM score <12
Table 12	CFA Goodness of Fit Indices for Participants with Impression Management scores less than 12
Table 13	Comparisons of ‘Summed Score’ CFA models with ‘Facet Score’ models
Table 14	Comparison of expected and observed factor loadings
Table 15	Invariance tests of Managers and Validity samples
Table 16	Big Five Mean Scores comparison between Field Study samples

Table 17	Mean Scores for the 2 Cluster Groups based IM, Agreeableness, Conscientiousness
Table 18	Frequency Table of IM scores in Total and in Cluster 2
Table 19	Cohen's 'd' Effect Sizes for Participants in Cluster 1 with IM Score of 12 or greater, compared with all other Participants.
Table 20	Correlations between the Predictor measures used in the assessments
Table 21	Simulation results of proportion of simulations containing Fakers, and the proportion of times a Faker is selected
Table 22	Effect of number of Job Applicants in Selection Set on proportion of simulations containing Fakers, and the proportion of times a Faker is selected
Table 23	Effect of Predictor Set on Rank Order of Job Applicants
Table 24	Effect of Different Cut-off Hurdles on proportion of simulations containing Fakers, and the proportion of times a Faker is selected
Table 25	Means and Standard Deviations of the bespoke BIDR-IM scale for different groups
Table 26	Effect sizes for Mean Score Differences in Impression Management scores between groups

Abstract

Fahey, Gerry (2017). *The Construct Validity of the NEO PI-R Personality Inventory in High Stakes Employee Selection.*

The purpose of this study was to establish the construct validity of the NEO PI-R personality measure when used for high stakes employee selection purposes. Based on extant research from industrial/organisational psychology, social psychology, and behavioural economics it is argued that deliberate impression management, or faking good, by job candidates in high stakes selection contexts can occur. This can be regarded as a form of moral hypocrisy. Theoretical research showed that moral hypocrisy occurs in ambiguous contexts in the absence of reminders of moral standards. It was hypothesised that the use of a formal warning would eliminate or minimise faking good by participants in a field study of job applicants in a high stakes contexts, thereby allowing construct valid inferences to be made about the participant's personality traits. To test this hypothesis a formal verbal warning about measures included in the assessment to detect deliberate impression management was given to the participants. They completed the NEO PI-R as part of the battery of tests used in the selection process for middle and senior management positions in a range of organisations. A bespoke impression management measure, based on a widely used measure used to detect deliberate impression management, was included in the battery of tests. A second field study sample was used to validate the findings of the managerial field study. Using confirmatory factor analysis the results showed that faking good was minimised, but not eliminated. Monte Carlo simulations showed that it was still possible that participants, who faked good in spite of the warning, could be selected from a short list of job applicants. The use of the bespoke impression management measure was shown to be of benefit in minimising bias and unfairness arising from the use of the personality measure when selecting a candidate from a short list.

QUOTE

“Let us make recommendations to ensure that NASA officials deal in a world of reality in understanding technological weaknesses and imperfections well enough to be actively trying to eliminate them” *Richard Feynman, Appendix to the Report on Challenger Disaster.*

Chapter 1

Introduction

Personality assessment using self-report measures is now a well established practice in the field of applied industrial/organisational, or occupational, psychology (Barrick, Mount, & Judge, 2001; Hogan, 2005; Hough & Oswald, 2008; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). The aim of the present research was to establish the construct validity of the NEO PI-R omnibus personality measure (McCrae & Costa, 2008) in a high stakes employee selection context. High stakes employee selection situations are those in which individual job candidates expect to benefit in some tangible or psychological way from a positive decision with respect to their job application.

The question of the accurate assessment of personality is an important theoretical question (Ellingson, 2012; Ployhart, Schmitt, & Tippins, 2017) with critical consequences in applied settings (Griffith & Converse, 2012). The present research contributes to theory on this question. From a theoretical perspective, the construct validity of self-report measures has long been a major question (Cronbach & Meehl, 1955; Kane, 2013). Accurate personality measurement has implications for understanding a fundamental question in personality psychology, namely, the hierarchical structure of personality (DeYoung, 2006; Chang, Connelly, & Geeza, 2012). From an applied perspective, bias and unfairness issues arise from inaccurate measurement of personality (Messick, 1995).

Objectively scored psychometric measures such as cognitive ability tests differ from personality measures which rely on self-reported data. An individual's test score has meaning on the basis that it provides a measureable link with the individual's behaviour in a particular setting of interest (Lance, Dawson, Birkelbach, & Hoffman, 2010). However, the question of whether individuals respond honestly to the items in personality measures, or whether they engage in what is usually referred to either as 'faking good' or 'impression management' is a hotly contested topic (Ellingson, 2012) with three opposing viewpoints on the topic. A number of researchers maintain that faking good is a serious problem (Griffith & Converse, 2012; Morgeson, Campion, Dipboye, Hollenbeck, Murphy, & Schmitt, 2007). Some researchers argue that even if faking good were to occur it would not matter because studies have shown that the occurrence does not affect the criterion related validity of self-report personality measures (Ones, Viswesvaran, & Reiss, 1996). Others argue that faking is not an issue in high stakes personality assessments (Hogan, Barrett, & Hogan, 2007). These opposing viewpoints raise issues pertaining to core psychometrics that question the construct validity of personality measures. This is so because they raise doubts about the accuracy of the inferences made about an individual based on her or his score on a personality measure in a personality assessment context particularly those that are described as 'high stakes'.

Arising from these questions of construct validity a programme of research was devised, using the NEO PI-R, in order to contribute to the debate by answering some of the questions raised in the preceding paragraph. The key issues that the research – literature review, field study and validation study, and Monte Carlo simulations – undertaken for this thesis was designed to address were:

1. How can the construct validity of a psychological instrument such as the NEO PI-R be properly established?
2. What is the theoretical and empirical evidence for the viewpoint that faking good does occur in personality assessments?
3. What is the evidence that its occurrence is detrimental to the construct validity of the NEO PI-R when used in high stakes employee selection situations?
4. What defensible methodological approach can be used to clearly demonstrate that the use of a formal warning concerning measures to detect faking as part of the assessment procedure was, or was not, effective in preventing faking good from occurring?
5. Is it possible, using a bespoke impression management measure, to detect those who faked good in spite of the warning?

Together these questions constitute the core of the research undertaken with a view to establishing the construct validity of the NEO PI-R when used for assessing an individual job candidate's personality traits in a high stakes employee selection context.

Based on the experimental psychology research findings of Batson and his colleagues (Batson, Kobryniewicz, Dinnerstein, Kempf, & Wilson, 1997; Batson Thompson, Seufferling, Whitney, & Strongman, 1999) participants in the present research were issued with a verbal warning, as a putative method for minimising the incidence of faking good, before completing an omnibus personality measure. It was hypothesised that following this procedure, participants would be less likely to engage in faking good and, as a consequence, their responses to items in the NEO PI-R would be more objective in the sense of more accurately representing the true self.

The research for this programme was carried out using participants from (a) a field study of job applicants for middle and senior management positions in a range of organisations, and (b) a validation sample of job applicants for senior positions in a large company. Participants in the managerial field study completed the well-known, and widely used, NEO PI-R omnibus personality measure. In addition, each participant completed a bespoke impression management measure, or lie scale, that was based on the widely used Paulhus's Balanced Inventory of Desirable Responding (Paulhus, 1984). Monte Carlo simulations were used to examine the effect of faking good on the extent of unfair selection decisions due to faking good by some job candidates.

The methodological technique of multitrait-multimethod (MTMM) investigations is recognised as the best statistical method to use in order to separate the effects of substantive traits from the contaminating effect on variance due to socially desirable responding so as to properly establish the construct validity of a trait when relying on self-report measures (Campbell & Fiske, 1959; Chang, Connelly, & Geeza, 2012). Both the field study and validation study of this research project were monomethod rather than multimethod studies. To, therefore, determine whether the use of a formal warning as part of the personality assessment was effective or not in eliminating or minimising the incidence of faking good, the methodological approach taken was to rely on an examination of the higher order structure of the Big Five dimensions of personality using confirmatory factor analysis (CFA). The CFA results from the field study sample were then compared with the findings of extant research on the topic of the higher order structure of personality using an MTMM methodology. If the results of the monomethod field studies, with respect to the Big Five higher order structure, are found to be consistent with the

findings of the MTMM studies this would confirm the findings concerning the higher order structure of the Big Five in a monomethod study because of the comparison with the more methodologically reliable findings of the MTMM studies (McDonald, 1999, pp. 213-222).

By adopting this methodological approach it was possible in the field studies to examine two fundamental issues. Firstly, whether or not the formal verbal warning used in the assessment of participants was effective in eliminating or minimising faking good. Secondly, if it was effective in this research objective then it would be possible, for the first time in a monomethod study, to shed light on the question - does the higher order structure of the Big Five dimensions of personality actually exist or is it simply a statistical artefact arising from the variance due to socially desirable responding? (DeYoung, 2006).

Single studies (Roberts, Kuncel, Shiner, Caspi, Goldberg, 2007) as well as meta-analytic study findings (Ones, Dilchert, Viswesvaran & Judge, 2007) support the usefulness of personality measures in predicting aspects of behaviour such as job performance, leadership, organisational citizenship behaviour, teamwork, interpersonal behaviours, and counterproductive work behaviour in work and organisational settings. Specifically, Ones et al. (2007) conclude that “any selection decision that does not take the key personality characteristics of the job applicants into account would be deficient” (p. 1020). According to Hogan (2005), personality predicts occupational performance almost as well as measures of cognitive ability. Unlike cognitive ability measures, personality measures do not discriminate against job candidates because of the aggregated group effects due to differential item functioning arising from individual differences in gender or race (Ones et al., 2007). Personality measurement for selection is now part of a ‘multibillion dollar

international industry' (Ziegler, MacCann, & Roberts, 2012, p. 3), and this research has immense practical implications for that industry.

Therefore the accurate assessment of employees benefits both the individual from a job fit perspective, and the organisation from a financial perspective. For example, employee recruitment and selection procedures as part of high performance work practices have been found to affect employee behaviour (i.e. withdrawal behaviour, productivity), which in turn impacted both short term and long term corporate financial performance (Huselid, 1995). Organisations that use personality measures in the selection and recruitment of managers, and which retain these employees, are likely to outperform their competitors that do not select on personality (Hogan, Hogan, & Kaiser, 2010; Oh, Kim, & Van Iddekinge, 2015). These benefits of personality assessment, however, only arise if the individual's true score on a personality measure is accurately assessed, thus being indicative of 'construct validity'.

In defending the use of personality measures in applied situations such as high stakes employee selection situations, Hogan (2005) stated that "the problem is that business people have trouble getting good advice from academic psychology. This in turn explains the widespread interest in bogus measures of personality such as the Myers–Briggs Type Indicator and Goleman's Emotional Competence Inventory" (p. 334). The construct validity of personality measures such as the NEO PI-R in high stakes employee selection situations is a fundamental measurement issue with respect to the inferences made about the personality traits of job candidates. If the observed score on such measures is not aligned with the individual's true score on a putative latent construct of each of the dimensions of personality then the observed score is not a valid measure, and is as open to the same criticism as that levelled by Hogan at the

Myers-Briggs Type Indicator or Goleman's Emotional Competence Inventory. This thesis seeks to evaluate a procedure for the administration of self report personality measures, such as the NEO PI-R, that results in construct valid inferences about job candidates' personalities.

Section 2.1 of the thesis starts with an in depth literature review in Chapter 2 of the theoretical research dealing with what exactly is meant by 'construct validity'. Chapter 3 then provides a review of relevant aspects of the body of research in support of the present day understanding of personality, with an emphasis on its hierarchical structure and role in understanding employee behaviour in the workplace. In Chapter 4 the thesis will cover socially desirable responding and its effect on the measurement of the true score of latent personality constructs. Research on the effect of variance due to socially desirable responding on the findings regarding the hierarchical structure of personality is also covered in this chapter. Chapter 5 contains a comprehensive literature review of research on the topics of 'faking' from the field of industrial/organisation psychology, as well as relevant research on the related topics of moral hypocrisy and moral disengagement from other research areas such as social psychology and behavioural economics. Investigations of faking good, arguably a behavioural manifestation of moral hypocrisy, by researchers in industrial/organisational psychology have, to date, largely ignored relevant research from these related fields on the topic of moral hypocrisy. In Chapter 6 there is a review of the three main methodological issues that impact on the analysis of the research findings of the field studies – the separation of trait effects from method effects, factor analysis considerations, and the ability of the impression management 'lie scale' used in the field study and simulations to detect faking good. Chapter 7 covers the methodological and analytical procedures followed to establish the

construct validity of the NEO PI-R for the field research and the Monte Carlo simulations that were carried out. Chapter 8 presents details of the analyses of the results of the field studies and simulations. Finally, Chapter 9 provides a discussion of the research findings and the conclusions arrived at.

Chapter 2

The Concept of Construct Validity

Validity remains the most important term in the educational and psychological measurement lexicon, according to recent research (Byrne, Peters, & Weston, 2016; Newton & Baird, 2016). This thesis is in essence about construct validity – how it is properly evaluated, and particularly how it relates to the accurate assessment of personality as well as the hierarchical structure of personality. The validity of a psychological test refers to the inferences that are made about the test score rather than simply being a property of the test itself (Cronbach & Meehl, 1955).

The present chapter therefore introduces and defines the concept of construct validity, and reviews the respective literature. Specifically, the chapter first gives an account of the fragmented approach (Strauss & Smith, 2009) to the topic of validity which was the dominant approach to the topic prior to the 1950's. This account is followed by a review of a small number of landmark articles that form the basis for the modern theoretically grounded, and unified, approach to the topic of construct validity (Cizek, Bowen, & Church, 2010; Cook, 2006; Embretson, 2009; Messick, 1989, 1995; Kane, 2001, 2013; Strauss & Strong, 2009). The importance of these articles is referred to by Strauss and Smith (2009) as follows that “Indeed, theoretical progress in clinical psychology has substantially depended on four seminal papers all published within a decade” (p. 6) in their review article on the topic of construct validity.

2.1 The pre 1955 Fragmented Model

Psychological tests can be used for two different purposes n measurement and prediction. A number of different approaches have evolved which have been used in deconstructing the validity of a psychological measure into its component parts. Usually, however, the two purposes are combined i.e. the measurement of a latent construct is used to make a prediction and/or offer an explanation of behaviour (Murphy & Davidshofer, 1998). There is an important difference between measurement in the physical and behavioural sciences. In the physical sciences objectively determined gold standard measures may be available for reference such as the metre stick in the International Bureau for Weights and Measures (Quinn, 1999). In contrast, in the behavioural sciences, the constructs are latent i.e. their existence is hypothesised and arrived at by an inductive process, and are measured by inference (McDonald, 1999). For example, a score on an IQ measure such as the Wechsler Adult Intelligence Scale Revised is taken to be a measure of an underlying property of the human mind that some would argue is very ill defined (Haier, Colom, Schroeder, Condon, Tang, Eaves, & Head, 2009). There is simply no easy way to determine whether or not a psychological measure validly reflects the construct. Both the researcher and practitioner in the behavioural sciences must always have some degree of healthy scepticism when it comes to the question of the construct validity of measures (Hogan, 2005).

Construct validity is frequently seen to include a number of what were regarded as essentially independent aspects of validity – content, criterion (concurrent and predictive), and construct (Kane, 2001, Strauss & Smith, 2009). Face validity is also sometimes included as being an additional aspect of construct validity (Murphy

& Davidshofer, 1998). The use of this latter term is discouraged by some researchers because an evaluation of the superficial qualities of psychometric measures is perceptual in nature (Cook, 2006).

Face validity represents an interaction between what the test asks the test takers to do and the test takers' understanding of what the test is designed to measure (Murphy & Davidshofer, 1998). If test takers' perception is that the test they are taking is not actually relevant to what is being assessed, then they may not respond accurately to the items in the test (Cook, 2006). For example, random responding and nay saying can occur in personality measures if the individual taking the test does not regard the test as a relevant measure (Costa & McCrae, 1995).

Content validity according to Murphy and Davidshofer (1998) refers to the aspect of validity that "is established by showing that the behaviours sampled by a test are a representative sample of the attribute being measured" (p.149). This perspective on validity has to do with the measurement use of a test. It requires that the items in a test measure each domain that the test covers (Clarke & Watson, 1995, Messick, 1995). For example, an intelligence or cognitive ability test should contain items that assess the verbal, numerical, and figural domains (Gottfredson, 1997). It is therefore logical that a measure of the Big Five dimensions of personality should contain items that systematically sample each of the five personality dimensions.

Criterion related validity compares test scores on a measure with test scores on some other attribute of interest (Ones & Viswesvaran, 1997). It is concerned with prediction, and has two aspects – concurrent and predictive. For example, in psychotherapy, a person's score on Beck's Depression Scale could be compared with an individual's score on Rotter's Locus of Control measure, with both measures assessed contemporaneously or essentially at the same time (Borckardt, Nash,

Murphy, Moore, Shaw, & O'Neil, 2008). This is an example of concurrent criterion validity in that two contemporaneous aspects of an individual's psychological state are compared. A prime example of predictive criterion validity can be seen in which a measure of an individual's IQ, such as the SAT score, is used to predict future performance e.g. job or academic related (Hogan, Hogan, & Roberts, 1996). This approach to validity becomes problematic when a criterion based approach is used to validate a psychological measure, for example in those situations where there might be no criterion available (Kane, 2001). For example, in some employee selection situations, there may be no easily measureable criterion for job performance available, such as that of the CEO of a large company in which the lead time between strategic initiatives and financial outcome is lengthy (Mlodinow, 2009).

This fragmented approach to validity can still to be found in the promotional literature of many test publishers as a number of researchers have pointed out (Goldberg, Johnson, Eber, Hogan, Ashton, Cloninger, & Gough, 2006; Hogan, 2005; Kane, 2001, 2013). As Cronbach (1980) pointed out, "The great run of test developers have treated construct validity as a wastebasket category" (p. 44).

2.2 The Unified Model

The definition of construct validity used in this research is that of Messick (1995), who defines construct validity as being "an overall evaluative judgment of the degree to which multiple forms of evidence and theoretical forms of rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores" (p. 741). This definition, as a general framework for establishing construct validity, reflects the modern approach to construct validity and owes much

to a small number of landmark papers - Cronbach and Meehl (1955), Loevinger (1957), and Campbell and Fiske (1959). It would be very unusual to find a review of the theoretical underpinnings of construct validity that did not refer to these three articles as well as Messick's (1989, 1995) more recent publications on the topic (Bornstein, 2011; Borsboom, Mellenbergh, & van Heerden, 2004; Cizek, 2012; Cook & Beckman, 2006; Downing, 2003; Goodwin & Leech, 2003; Kane, 2001, 2013; Strauss & Smith, 2009).

The fragmented aspects approach to validity represents the approach that prevailed prior to the 1950's in psychology (Kane, 2013) and can be contrasted with the unified approach to validity, the topic of this section of the chapter. The unified view of construct validity, that Messick's definition encapsulates, relies on a theoretical approach to the topic which has emerged from the three landmark papers (Kane, 2001). Therefore, a review of a number of the seminal (Cook, 2006; Kane, 2013; Strauss & Smith, 2009) articles dealing with the theoretical underpinnings of the concept of construct validity, is a necessary precondition for gaining an understanding of what the scientific concept of the process of construct validation, which is central to this research programme, entails.

2.2.1 Cronbach and Meehl on Construct Validity

Cronbach and Meehl (1955, p. 282) made the important theoretical point that the process of construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality even when it is not 'operationally defined'. It is often studied when the tester has no definite criterion measure of the attribute of interest, and must use indirect measures. This definition of construct validity, as

articulated by Cronbach and Meehl (1955), is much broader in scope than that of content and criterion related validity. They saw validation as an inductive process based on multiple sources of evidence including content and criterion related evidence. Therefore, the problem faced by the investigator, is the question “what constructs account for variance in test performance?”. The term first used for this aspect of validity was ‘congruent validity’. This was later changed to ‘construct validity’ by an American Psychological Association (APA) Committee on Psychological Tests when, following the publication of the Cronbach and Meehl (1955) article, the APA undertook to specify what qualities should be investigated before a test is published.

At that time the issue of test validation had become a major theoretical concern because prior to Cronbach and Meehl’s (1955) article on the topic validity was assessed by examining the content and criterion related validity of a test. Cronbach and Meehl’s (1955) critique of the extant approach to validity is consistent with Popper’s (1963) view of the method of scientific inquiry:

“Criticism of our conjectures is of decisive importance: by bringing out our mistakes it makes us understand the difficulties of the problem which we are trying to solve. This is how we become better acquainted with our problem, and able to propose more mature solutions: the very refutation of a theory - that is, of any serious tentative solution to our problem - is always a step forward that takes us nearer to the truth. And this is how we can learn from our mistakes”. (p. vii)

The Cronbach and Meehl (1995) paper also introduced the now important construct validation concept of the ‘nomological network’ surrounding a focal construct, which they defined as “the interlocking system of laws which constitute a theory”. Prior to Cronbach and Meehl’s (1995) paper validity was de facto a fixed or static concept based on content and criterion related validity (Kane, 2001; Strauss & Strong, 2009). The introduction of the concept of the nomological network underlying a psychological construct introduced the concept of learning more about the construct as further construct validation research occurred. Interestingly in their six statements of the principles, or guide rules, of a ‘nomological net’ Cronbach and Meehl (1995) use the word ‘nomological’ as a noun without ever actually defining the meaning of the noun, while today it is mainly used as an adjective.

There is a correspondence between ‘laws’ and ‘nomologicals’ in the 1955 paper but the latter term is, strictly speaking, broader than normal definition of a scientific law. To quote Cronbach and Meehl (1955), “The laws in a nomological network may relate (a) observable properties or quantities to each other; or (b) theoretical constructs to observables; or (c) different theoretical constructs to one another. These ‘laws’ may be statistical or deterministic” (p. 290). The most important aspect of the paper, and its propositions, is that it allows for a certain degree of fuzziness in the understanding of a construct, unlike the more definitive approach of both content and criterion related validity measures. This, the authors claim is due to the fact that “Psychology works with crude, half explicit formulations” because construct validation is an inductive rather than deductive process.

In conclusion, Cronbach and Meehl (1995) listed eight points concerning construct validity that they regard as ‘particularly significant’. The most important of these, from the perspective of this thesis, contains the comment “Many types of

evidence are relevant to construct validity, including content validity, inter-item correlations, inter-test correlations, test-‘criterion’ correlations, studies of stability over time, and stability under experimental intervention”. A construct to be valid must be validated across a range of ‘nomologicals’ in its nomological network. In the final paragraph the authors seem to be setting their concept of construct validation against the then existing practice of validation based simply on an ‘operational’ approach of content and criterion related approaches to the issue. According to the authors, this operational approach “would force research into a mould in which it does not fit” (p. 300).

Cronbach and Meehl’s (1955) paper represented a watershed in the approach to validation in that it added construct validity to content and criterion related validity as another separate and distinct aspect to be considered. It also represented a theoretical shift from the perspective of viewing validity as a property of the test to the now universally theoretically accepted perspective of viewing validity as an inference arising from test use. Cronbach and Meehl’s (1955) was expanded upon by Loevinger (1957) in her landmark monograph which is reviewed in the next subsection.

2.2.2 Loevinger on Construct Validity

The Cronbach and Meehl (1955) paper was followed shortly by Loevinger’s (1957) monograph. Loevinger advanced the concept of construct validity beyond what Cronbach and Meehl posited and began the process of unifying the different aspects of validity evidence. In the criterion model of validity, the test scores are simply compared to the criterion scores. In the content model, the characteristics of the

measurement procedure are evaluated in terms of expert opinion about how the observable variable should be measured. In the construct validity model, the evaluation of validity always requires an extended analysis. As a result, the development of the construct validity model highlighted the inadequacies of most validation efforts based on a single (often dubious) validity coefficient or simply an expert opinion (Kane, 2001). By unifying these different aspects of validity evidence (Clarke & Watson, 1995) Loevinger's monograph is the "most complete exposition of theoretically based psychological test construction" (p. 308).

Instead of viewing constructs as measures of attributes which are not 'operationally defined' Loevinger felt that Cronbach and Meehl were being too reluctant to assign reality to constructs or traits in their definition of what the concept of construct validity was. She compared the relationship between a trait and a construct as analogous to the distinction and relationship between a parameter and its corresponding statistic. Construct validity concerns the validity of the inference made about the use of a test as a measure of a trait which existed prior to, and independent of, the psychologist's act of measuring. The trait is what psychologist aim to understand and the construct is the *current* best understanding of the trait. This implies, as Cronbach and Meehl also did to a lesser extent, that construct validation is a dynamic process involving induction and a nomological network.

The monograph of Loevinger introduced three aspects of construct validity, namely, the substantive, structural, and external components. The substantive component encompasses content validity, but is broader than it and is concerned with how best to delineate the construct domain or domains of interest. The key substantive issue to be resolved in the initial developmental stage is the scope or generality of the target construct. Clarke and Watson (1995) provide a very readable 'exegesis' of that

part of Loevinger's monograph dealing with the substantive component. Their article provides practical guidance for applying Loevinger's theoretical approach to the practical problem of actually developing a psychometrically sound measure. This doctoral thesis is not concerned with this substantive component in that participants in the field studies completed personality measures, the NEO PI-R or NEO-PI3, which have already been subjected to the developmental process of test construction based on well established psychometric principles. The NEO is a widely used , extensively researched, Big Five personality measure. The focus of the thesis is primarily concerned with the substantive and external components of construct validity as defined by Loevinger.

The structural component of construct validity is concerned with item selection and the psychometric evaluation of the test with respect to the homogeneity or unidimensionality of the psychometric measure of a putative construct using a technique such as factor analysis (FA) or item response theory (IRT). Loevinger's monograph did not have the benefit of the accumulated body of knowledge and research concerning FA and IRT that Clarke and Watson (1995) had. Essentially the structural component deals with the techniques of item selection and, today, the technique of confirmatory factor analysis (CFA) achieves what Loevinger based her concept of the structural component of construct validity on for tests which are scored using a Likert type scale. Item Response Theory (IRT) achieves a similar objective for tests which use dichotomous scoring (McDonald, 1999).

The treatment of the external component by Loevinger is consistent with the widely used division between concurrent and predictive criterion related validity. Discriminant validity was shown to be of importance with respect to the external component of construct validity. Finally, the term 'distortions of measurement' was

included in the monograph. These are errors of measurement which are correlated with true scores which are not random, because randomness is a fundamental assumption of classical test theory. As far as Loevinger was concerned demonstrations of negligible relationships with known sources of distortion is an *essential* rather than optional step in test validation. This is a critical point with respect to the subject matter of this thesis, the methodology of which is largely concerned with the impact of the distorting effect of common method variance (CMV) on the measurement of personality traits. Brown (2006) states that “In sum, construct validation is limited in instances where a single assessment method is employed” (p. 214). The research approach used for this thesis was based on a field study and validation sample both of which used a single assessment method. Therefore, methodological consideration had to be given to the issues that arose from relying on the monomethod studies used in the research. The solution to this problem is dealt with detail in Chapter 6.

2.2.3 Campbell and Fiske and the Multitrait-Multimethod Approach

The third landmark paper concerning construct validity is Campbell and Fiske’s 1959 paper dealing with the ‘multitrait-multimethod’ approach to construct validation and this landmark paper is relied upon indirectly, in the empirical research approach taken in this thesis, to deal with the issue of method variance. It is narrower in scope than the approaches of either Cronbach and Meehl or Loevinger in that it is not concerned with content validity or Loevinger’s substantive and structural component of validity. It is primarily concerned with a test’s correlational relationship

with other tests. Campbell and Fiske (1959) recognised that the statistical phenomenon of shared method variance i.e. variance due to the common assessment method used such as self and peer reports, could account for substantial overlap among psychological measures. Because of the ever present, often substantial method variance in all psychological measures, the multitrait-multimethod approach of Campbell and Fiske to validation requires the simultaneous consideration of two or more traits measured by at least two different methods (McDonald, 1999).

If a number of methods, such as self, spouse, and employer evaluations of personality, are used to assess an individual's personality Big Five traits these measures will take the form of a multitrait-multimethod matrix. The size of the correlations between the methods for each trait will indicate the level of convergent or discriminant validity of the methods used. An additional benefit of using Campbell and Fiske's methodology is that the pattern of correlations between different constructs using the same method, when compared with the correlations between different traits measured by different methods, can be used to indicate the presence or absence of CMV (Brown, 2006; McDonald 1999; Murphy & Davidshoffer, 1998). Evidence of convergent and discriminant validity with other tests is an atheoretical operational approach to establishing an external aspect of construct validity (Campbell & Fiske, 1959). The importance of the Campbell and Fiske paper lies in the fact that it can be used to investigate the impact of CMV on the measurement of a construct. This is very important when it comes to establishing the construct validity of self-report measures of observed indicators of constructs and their corresponding latent traits. Since the publication in 1959 of Campbell and Fiske's paper there have been major advances in methodological techniques. These include factor analysis and structural equation modelling for evaluating multitrait-multimethod correlation

matrices, which have made the quantification of convergent and discriminant validity as well as the variance due to method variance much more accurate (Brown, 2006). Ozer (1999) emphasised that successful construct validation of personality assessment inferences requires the use of both mono- and multi-method approaches. This is the basis for the methodological approach adopted in the research programme of this thesis using the field studies as well as research findings from extant MTMM studies.

2.2.4 Messick's and Embretson's Contributions

A number of other theorists have helped to develop a modern general framework for the unified model of construct validity arising from the seminal papers described above (Kane, 2013). This was due to further conceptual and empirical evaluation, based on developments in statistical techniques, of the structural aspects of construct validation. Foremost among these theorists were Messick and Embretson whose respective contributions to the modern concept of construct validity are next reviewed.

In 1989 Messick published a paper entitled 'Meaning and values in test validation: The science and ethics of assessment' and a similar paper in the *American Psychologist* in 1995 (Messick, 1995). Messick's theoretical approach to construct validity is built on the Cronbach and Meehl (1955) and Loevinger (1957) papers, as well as that of Campbell and Fiske (1959). Messick's definition of construct validity has become the accepted benchmark against which psychological tests must be evaluated (Kane, 2013). His 1995 paper is subtitled 'Validity of inferences from persons' responses and performances as scientific inquiry into score meaning'. This title suggests that it is the inference taken from the score on a test that is important

from a validity perspective, and not the score per se. According to Messick, construct validity is a unified concept which integrates all sources of validity information from the nomological network into an integrative summary of the meaning and consequences of a test score. It encompasses content and criterion related validity measures as well as convergent and discriminant evidence of validity. It also includes Loevinger's substantive, structural, and external components. It gives as much weight to the external consequences of test use as previously given to in the 'internal' aspects of developing a test – hence the focus on 'values' in the subtitle of Messick's (1995) paper in the *American Psychologist*.

According to Messick's 1989 and 1995 (p. 745) papers, a unified concept of construct validity contains six sources of evidence:

1. The content aspect of construct validity includes evidence of content relevance, representativeness, and technical quality
2. The substantive aspect refers to theoretical rationales for the observed consistencies in test responses, including process models of task performance, along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks
3. The structural aspect appraises the fidelity of the scoring structure to the structure of the construct domain at issue
4. The generalisability aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks, including validity generalization of test criterion relationships
5. The external aspect includes convergent and discriminant evidence from multitrait-multimethod comparisons, as well as evidence of criterion relevance and applied utility

6. The consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice.

The importance of Messick's unified concept of construct validity is that each application of a measure should be evaluated on its own merits and, secondly, that the consequences of test use are an integral part of construct validation. One consequence of this, according to Kane (2001), is that in high stakes selection situations issues such as CMV and job candidate coaching for test taking can easily impact on the validity of the inferences drawn from test scores.

At the same time as Messick was contributing to the debate on construct validity Embretson (1983) put forward the argument that, since Cronbach and Meehl's articulation of the concept of construct validity, research into construct validation consists essentially of two stands, namely, 'construct representation' and 'nomothetic span'. *Construct representations* are concerned with identifying the theoretical mechanisms that underlie item responses, such as information processes, strategies, and knowledge stores. *Nomothetic span* is concerned with the network of relationships of a test score with other variables. These two types of construct validation research address different issues, and require different types of data. She further elaborated on what this research strategy into the validity of a construct entailed (Embretson, 2007).

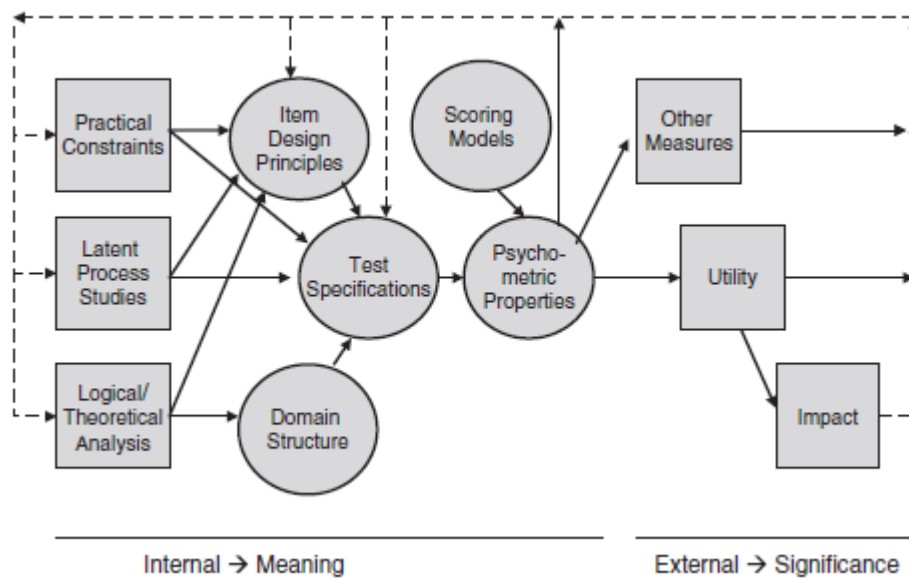


Figure 1 *Embretson's Universal System for Construct Validity*

This model, in Figure 1 above, of Embretson captures all the elements of the theoretical views of Cronbach and Meehl, Loevinger, and Messick concerning the modern concept of construct validity. Today there is a greater appreciation for the indeterminate and ongoing nature of theory building, and the need for theory revision as a result of scientific criticism arising from continuing research (Kane, 2013). In particular Embretson describes her model as a universal system for evaluating validity, and it captures the dynamic and on-going nature of the validation process. It is also interactive because all of the elements in the system have an effect on, or are affected by, other elements in the nomological network – solid lines in Figure 1 indicate direct paths of interaction, dashed lines indicate feedback paths. The lines at the bottom of Figure 1 are meant to delineate the substantive and structural aspects of the validation process from the external aspects including social consequences. Embretson's universal model is also consistent with Cizek's (2012) recent definition of validity:

“Validation is the ongoing process of gathering, summarizing, and evaluating relevant evidence concerning the degree to which that evidence supports the intended meaning of scores yielded by an instrument and inferences about standing on the characteristic it was designed to measure” (p. 35).

From the perspective of Embretson’s universal system, construct validation is primarily concerned with the following specific elements of the system of validation evidence categories:

1. Logical/Theoretical Analysis – Theory of the subject matter content, specification of areas of interest and their interrelationships
2. Latent Process Studies – this includes the impact of testing conditions and test administration methods on participants’ responses to items
3. Test Specifications – the specification of testing conditions forms part of this evidence category
4. Utility – relationship of scores to external variables, criteria, and categories
5. Impact – consequences of test use, adverse impact.

This categorisation is very similar to that of the six aspects approach of Messick (1995), and adds to it by giving greater emphasis to construct validation being a dynamic process with feedback effects that can lead to a review of prior theory. It should be noted here that the pre 1955 view of validity has some adherents who still maintain that validity is simply a property of the test (Borsboom, Mellenbergh, & van Heerden, 2004). This view of validity is rejected by Kane (2013) on the basis that the Borsboom et al. (2004) argument depends on both the specification and evaluation of

a causal property that exists independent of the researcher. This also requires a well established theory of test performance that specifies how the property produces the observed test performance. There is, as yet, no published research which supports this position (Hughes, in press).

To summarise Section 2.2 the theoretical views of Cronbach and Meehl (1995), Loevinger (1957), and Campbell and Fiske (1959), as combined in their models by both Messick (1995) and Embretson (2007), support the position that the establishment of construct validity must take account of a number of different perspectives and sources of evidence. The importance of this approach for the research programme is reviewed in the next section.

2.3 The Importance of Construct Validation

Validity essentially concerns the meaning of inferences concerning test scores. It is best viewed as a carefully structured argument in which evidence is assembled in support of or to refute proposed interpretations of results from multiple sources (Cook, 2006). A failure in carrying out personality assessment to ensure that the requirements for achieving construct validity are met means that CMV arising from socially desirable responding may be a major concern. This is because of the reliance on self-report measures of the Big Five dimensions of personality in high stakes selection situations. If a job candidate is deliberately not responding honestly to the items in the NEO PI-R then Embretson's external evidence categories of Utility and Impact are immediately affected. This may result in potential negative consequences for all candidates who are being considered for employment, as well as the

organisation seeking to fill a position. This in turn leads to questioning of the first three evidential categories of Embretson listed at the end of the previous section. Questions concerning test administration and the specification of testing conditions become relevant, as well as how best to ensure accurate measurement of the Big Five dimensions of personality in high stakes employee selection situations.

Consistent with the Section 2.1 and 2.2 approaches to construct validity Cronbach made an important distinction between ‘strong’ and ‘weak’ construct validity (Kane, 2001). He regarded weak validity as an approach that relies on any evidence even remotely connected to the test scores as exemplified by the fragmented ‘aspects’ approach described in Section 2.1. At the applied level, where the promotional material of test developers is frequently relied upon by practitioners, the shortcomings of the weak approach to construct validity become a real issue. Cronbach (quoted in Kane, 2001; p. 326) had this to say, “The great run of test developers have treated construct validity as a wastebasket category. In a test manual the section with that heading is likely to be an unordered array of correlations with miscellaneous other tests and demographic variables. Some of these facts bear on construct validity, but a coordinated argument is missing”.

On the other hand, in essence, the strong approach to the establishment of construct validity of a psychological test properly refers to the inferences that are made about the test score rather than simply being a property of the test itself (Cronbach & Meehl, 1955). This is a crucial point because in practice test developers and others frequently, using the weak approach, erroneously refer to the ‘validity of the test’ as if validity is an inherent property of the test itself rather than the inference made from the test score (Clarke & Watson, 1995; Kane, 2013, Smith, 2005; Strauss & Smith, 2009). Hogan (2005) addressed the consequences of this issue when he

expressed the view that one of the main problems with personality psychology is “a generalised lack of concern for measurement validity” (p. 332). He reckoned that of an estimated 2,500 test publishers in the United States very few pay serious attention to the validity of the tests that they publish, and that publishers generally market tests with no demonstrated validity. He instanced two widely used exemplars of this ignoring of measurement validity, namely, the Myers Briggs Typology Indicator (MBTI) and Goleman’s Emotional Competence Inventory. The scepticism to be found surrounding the topic of personality assessment (Paul, 2004) is due entirely to the willingness of test publishers to disregard construct validity, in Hogan’s view. The disparity between test publishers’ approach to validity and that of theorists in the field partly explains why HR professionals are poorly informed of best practice in many areas (Hughes & Batley, in press; Rynes, Colbert, & Brown, 2002). Rynes et al. (2002) found, for instance that 51% of HR practitioners surveyed believed that there are really only four basic dimensions of personality as captured by the MBTI. Yet the MBTI has been shown by many researchers to be severely lacking in construct validity (Arnau, Green, Rosen, Gleaves, & Melancon, 2003; Bess & Harvey, 2002; Pittenger, 1993; Pittenger, 2005; Saggino & Kline, 1996; Saggino, Cooper, & Kline, 2001).

The importance of a theoretically based understanding of the concept of construct validity for this research programme can be seen from the debate in the literature concerning the effect of impression management on the validity of personality assessment. This is due to impression management in the form of faking good or the intentional distortion of responses to items in the self report personality measure used (Sackett, 2012). Ones and Viswesvaran (1998) conducted a meta-analytic study of the effect of social desirability on personality assessment for

personnel selection with a focus on criterion related validity. They state that “real world data show that social desirability is not a factor destroying the criterion related validity of personality measures” (p. 266). However, when it comes to the issue of the use of personality measures for job selection purposes Griffith and Converse (2011) present what they regard as compelling evidence that ‘roughly 30% of applicants are engaged in faking behaviour’ (p. 47). Both viewpoints are possibly correct simply because they deal with different aspects of construct validity, even though they might appear to be contradictory at a very superficial level of analysis. Sackett (2012) points out that because of socially desirable responding due to faking good there can be negative consequences. He defines faking good (p.331) as ‘situationally specific intention distortion’ in the context of job candidates responding to items in a self report measure in an employee selection context. Where top down selection is used to select the best candidate an unfairness problem can arise. This is because of the displacement of some candidates by those who engaged in faking good their responses to items in the measure used. Yet the criterion related validity of personality measures may well be unaffected by the problems with such measures that arise when used for job selection purposes (Hollenbeck, 2009).

The preceding paragraph vividly demonstrates the difference between a weak reliance on criterion related validity and the strong approach of theorists, such as Cronbach and Meehl, Loevinger, Campbell and Fiske, Messick, and Embretson, to construct validity with its inclusion of the social consequences as an integral component of construct validity. This, in essence, is at the heart of the subject matter of this thesis. Clearly if the effect of social desirability due to faking good in employee selection results in unfairness or bias in the selection decision, then the validity of the inferences about self-report personality measure with respect to the best

candidate are in error. Therefore, in that instance, a personality measure such as the NEO PI-R would be lacking construct validity even though measures of the Big Five such as the NEO PI-R do have criterion related validity when it comes to job performance prediction and evaluation. This is a very important applied problem when viewed from the perspective of high performance work practices (Huselid, 1995). As mentioned earlier an important element of high performance work practices is the use of best practice in the recruitment and selection aspect of the staffing practices.

The behavioural and social sciences in general have a somewhat questionable record when it comes to methodological rigour (Iaonnidis, 2005), and construct validity is no exception. In a recent edition of the journal *Assessment in Education: Principles, Policy & Practice*, devoted to the topic of validity, the editors (Newton & Baird, 2016) made the following comment that “Finally, it is important to recall Gafni’s observation that validation practice is often far from adequate and sometimes simply not conducted at all. We must not lose sight of the fact that there is far more to ensuring good validation than can be achieved by rigorous, scholarly debate over the meaning of validity” (p. 177). The second sentence could be taken to embrace the viewpoint of Hogan (2005), when he highlighted test publishers’ marketing of their proprietary tests, that there is often a lack of proper construct validation. This has led to in part to the scientist/practitioner gap in knowledge (Rynes et al. 2002). The most appropriate way to use personality measures for employee selection purposes is to first build a fully accurate measure of personality with many facets (Hughes, in press). Without this first step construct validity will never be established. The research programme of this thesis was designed to avoid this fundamental error.

Cizek et al. (2010) found in a survey of literature that the consequences of testing as a source of validity evidence is essentially nonexistent in the professional literature, and applied measurement and policy work. They referred to earlier research

which showed that of 238 tests in the Mental Measurements Yearbook, concurrent, and content validity evidence were provided fairly frequently (in 50.9%, and 48.4% of the tests, respectively), whereas evidence based on test consequences was noted for only two tests (0.7%). Cizek et al.'s own research found that, while out of 2,408 published articles 1,007 (41.8%) touched on validity, not one provided information related to consequences of testing as a source of validity evidence. As Cizek (2012) points out, "The usefulness of the score does depend, however, on the various contexts, decisions, or situations in which the test is applied. This is a separate effort in which empirical evidence and logical rationales must be conducted to determine if the (validated) score meaning is relevant to and justifiably used in the service of the diverse applications to which the test will be put" (p. 41).

This chapter reviewed the topic of construct validity in detail consistent with Cizek et al. (2010)'s call that increased attention be paid by researchers both to assuring confidence in the meaning of test scores and to investigating the consequences of test use. The issue of construct validity is critical for personality assessment, particularly in high stakes situations. Reliance on personality assessment procedures that do not take account of, and cater for, impression management is highly questionable from a construct validity perspective. If the procedures used by some organisations in personality assessment at the recruitment stage are questionable then it will not be possible for those organisations to meet the criteria for high performance work practices as defined by Huselid (1995). The research programme that was carried out to examine the construct validity of the NEO PI-R in such high stakes situations followed the unified model of construct validity. It focused on the big five personality factors which will be reviewed, along with the broader personality literature, in the next Chapter.

Chapter 3

The Dimensions of Personality

This chapter explores a number of aspects of the latent constructs of primary interest in this research project, namely, the personality traits or dimensions which have become known as the ‘Big Five’. The approach taken derives from Cronbach and Meehl’s (1955) exhortation that “A rigorous (though perhaps probabilistic) chain of inference is required to establish a test as a measure of a construct” (p. 291).

First, a broad overview of the current understanding of the Big Five dimensions is provided in Section 3.1 of this chapter. In Section 3.2 a review of the empirical evidence for the role of personality in determining workplace behaviour is provided. This sets the context for a comprehensive construct validation of the chain of inferences arising from the use of the NEO PI-R in high stakes employee selection contexts to be carried out, as both Embretson (2007) and Messick (1995) advocate. Sections 3.3 and 3.4 then provide a detailed review of the different putative models of the higher order structure of personality superordinate to the Big Five, as well as the putative aspects and facets of lower order structure that go to make up the Big Five dimensions. In Section 3.5, the empirical evidence in support of the putative higher order models is then reviewed in keeping with the need for elaborating on the nomological net (Cronbach & Meehl, 1955) requirement for establishing construct validity, as well as some measurement issues arising which are covered in Section 3.6.

3.1 The Big Five Dimensions of Personality

The descriptor 'Big Five' has today become synonymous with the topic of personality (Hogan, 2005). This model has its origins in a factor analytic research approach to understanding personality (Digman, 1990). The Big Five dimensions or, alternatively, the Five Factor Models (FFM) of personality, have achieved fairly widespread acceptance as a satisfactory explanatory model of the structure of personality and individual differences (Barrick, Mount, & Judge, 2001). Nevertheless, the debate is still on-going as to whether there are five factors or more at that domain level of analysis e.g. the Hexaco model (Ashton et al., 2004). The five broad dimensions or factors of Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness have been shown to be replicable across different demographic, ethnic, and cultural groupings (McCrae & Terracciano, 2005; Salgado, Moscovio, & Lado, 2003).

Based on descriptions given by Carver and Connor Smith (2010) the most prominent characteristics of the five dimensions can be summarised as follows: **Neuroticism** identifies the extent to which individuals are prone to experiencing psychological distress, and indicates lack of adjustment versus emotional stability. Individuals who score high on neuroticism are characterised by high levels of anxiety, hostility, depression, and self-consciousness. **Extraversion** identifies the quantity and intensity of energy directed by individuals outwards into the social world. High levels of extraversion indicate sociability, warmth, assertiveness, and activity, whereas individuals low on extraversion are described as reserved, sober, aloof, task-oriented, and introverted. **Openness to Experience** is defined as the active seeking and appreciation of experiences for their own sake. Openness to experience is defined in

terms of curiosity and the tendency for seeking and appreciating new experiences and novel ideas. Individuals who score low on openness are characterised as conventional, low in artistic engagement, and narrow in their range of interests. **Agreeableness** refers to the kinds of interactions with others an individual prefers ranging from being driven by compassion through to tough mindedness. Agreeableness is concerned with an individual's interpersonal orientation. It ranges from soft-hearted, good-natured, trusting, and gullible at one extreme to cynical, rude, suspicious, and manipulative at the other. **Conscientiousness** is a measure of the degree of organisation, persistence, control and motivation in goal directed behaviour that an individual possesses. Conscientiousness indicates the individual's degree of organisation, persistence, and motivation in goal-directed behaviour. Achievement-orientation and dependability have been found to be primary facets of conscientiousness.

These broad dimensions are proposed by many psychologists to be key personality determinants of behaviour, and the aggregation of information resulting from a person's standing on these dimensions gives a reasonably good assessment of what that person is like (Markon, Krueger, & Watson, 2005). There is also a growing body of evidence for a biological underpinning of the Big Five model of personality (Depue & Collins, 1999; Roberts & Jackson, 2008). The five dimensions or factors have been shown to have concurrent and predictive criterion validity in a number of different settings, such as occupational and clinical contexts (Barrick, Mount, & Judge, 2001; Caspi, Roberts, & Shiner, 2005; Samuel & Widiger, 2008; Roberts, 2009).

The 'five personality dimensions' model owes its existence to the work of Tupes and Crystal (1961) in the 1950's (Digman, 1990). They re-examined the data of Cattell, which led to his sixteen factor model of personality, known as the 16PF

(Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001) model of personality. They found that there were only five factors when Cattell's original data was factor analysed again. Subsequent to this research both the 'lexical' stream (the Big Five approach) the factor analysis of responses to omnibus personality questionnaires (the Five Factor Model approach) 'confirmed' the five factors by replicating Tupes and Crystal's findings. Digman, in particular, played a major role in establishing the validity of the five factor model of personality (Digman & Takemoto-Chock, 1981). The five factors are organised hierarchically in that the five broad dimensions subsume a number of facets. From a content validity perspective, the Big Five are composed of these narrower facet traits, which in turn comprise specific cognitive, behavioural, and emotional responses (McCrae & Costa, 1999). Different facets within a factor represent more specific characteristics, but the covariation of the facets indicates shared variance associated with a meaningful and more general underlying personality characteristic (Chang, Connelly, & Geeza, 2012). For instance, each of the broad dimensions of the NEO PI-R has six facets. However, each omnibus measure of the Big Five that is used has a different, although somewhat related, subordinate facet structure that makes up the Big Five (Hopwood & Donnellan, 2010). These measures (Salgado, 2003) include the NEO PI-R, Hogan's HPI (Hogan & Hogan, 1995), and the Personal Characteristics Inventory (Mount, Barrick, & Callans, 1995),

3.2 Personality and Behaviour in the Workplace

Personality affects work experience and, according to Roberts, Caspi, and Moffitt (2003), "Work experiences may alter personality; they make us more of who we already are" (p. 592). So an understanding of this is an important aspect of the

nomological net (Cronbach & Meehl, 1955) of personality as it relates to the domain of job performance, which involves both the external and consequential aspects of construct validity that Messick (1995) highlights.

It is also a very broad topic and will not be covered in great detail here apart from research linking the five personality dimensions to job performance. The importance of accurate personality assessment from a HR perspective lies in its value in aiding a better understanding of human behaviour in work and organisational settings both at an individual and aggregated level. Personality predicts job performance, but is not the only predictor (Hogan, Hogan, & Roberts, 1996). A meta-analysis of personality and overall assessment centre ratings (OAR's) conducted by Collins, Schmidt, Sanchez-Ku, Thomas, McDaniel, and Le (2003) found that, although cognitive ability alone predicted much of the variance in OAR's, the addition of personality traits to the model significantly increased the variance accounted for. In fact, they showed that in certain contexts, the combination of a set of personality traits and cognitive ability can predict nearly all of the variance in performance ratings.

Unlike cognitive ability, or intelligence in the vernacular, it is only since the early 1990's that the dimensions of personality have been seen to be of value when it comes to work and organisational settings (Barrick, Mount, & Judge, 2001; Ones, Dilchert, Viswesvaran, & Judge, 2007; Salgado, 2005). Research into the role of personality and its relationship to job performance suffered for quite some time from criticism by social psychologists, such as Walter Mischel in particular, who argued for a very 'situationist' approach to understanding human behaviour (Hogan, 2004; Roberts & Caspi 2001; Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). It was mainly due to a number of landmark meta-analytic studies that the emphasis changed

from a situationalist perspective to one based on relatively enduring personality traits. These studies carried out in the early 1990's showed that criterion related validity between the Big Five dimensions and job performance (Barrick & Mount, 1991; Hough, 1992; Tett, Rothstein, & Jackson, 1991) was significant and meaningful. This provided the first robust support for the use of measures of the Big Five in organisational settings. A large number of primary studies, as well as several meta-analyses conducted and published since the early 1990's, have provided on-going evidence for using personality measures in staffing decisions (Ones et al., 2007). As a result of this body of research the role of personality dimensions in work and organisational settings received a major boost.

Meta-analytic evidence (Barrick & Mount, 1991; Salgado, 2003) suggests that some of the Big Five dimensions are related to overall job performance in virtually all jobs, whereas other dimensions are related to performance in a more limited number of jobs. Conscientiousness has been empirically shown to be a valid predictor of job performance across performance measures in all occupations studied (Salgado, 2003). Neuroticism has also been found to be a generalisable predictor when overall work performance was the criterion, but its relationship to specific performance criteria and occupations was less consistent than was conscientiousness (Barrick, Mount, & Judge, 2001). Extraversion has been found to be related to job performance in occupations where interactions with others form a significant portion of the job such as jobs in the sales and marketing area (Barrick et al., 2001). Agreeableness is a useful predictor of service orientation and teamwork, because it has been demonstrated to have high predictive validity in jobs and work settings that involve considerable interpersonal interaction. This is particularly true when the interaction involves helping, cooperating with and nurturing others (Mount, Barrick, & Stewart, 1998).

Extraversion and Openness to experience appear to be related to training proficiency and creativity (Barrick et al., 2001, Salgado, 2005).

Ones, Dilchert, Viswesvaran, and Judge (2007) conducted a detailed review of the most comprehensive meta-analyses that have examined the relationships between the Big Five and a number of work related variables. These include (a) performance criteria (e.g., overall job performance, objective and task performance, contextual performance, and avoidance of counterproductive behaviours), (b) leadership criteria (emergence, effectiveness, and transformational leadership), (c) other criteria such as team performance and entrepreneurship, and (d) work motivation and attitudes. They showed that the accumulated body of evidence proves that the criterion-related validities of personality measures are substantial. The Big Five personality variables as a set do indeed predict important organisational behaviours such as job performance, leadership, and even work attitudes and motivation. The effect sizes for most of these criteria are moderate to strong (Salgado, 2005). Judge, Bono, Ilies, and Gerhardt (2002) have shown that leadership was related to the Big Five dimensions of personality. Extraversion was found to be the most important trait of leaders and effective leadership. After Extraversion, the dimensions of Conscientiousness and Openness to Experience were the strongest and most consistent correlates of leadership.

Not all personality traits are created equal in terms of their predictive and explanatory value (Markon et al., 2005). As a result, the highest criterion related validities for predicting overall job performance using predictors from the personality domain are found for compound personality variables (Ones, Viswesvaran, & Dilchert, 2005). Compound personality measures have been shown to have criterion related validities that are equal to those of cognitive ability, the best single predictor

of job performance (Schmidt & Hunter, 1997). Measuring personality traits is not simple. Items in personality measures can capture trait variance from cross loadings on factors and the facets of the Big Five, due to both the primary and secondary loadings of items in a personality measure (Johnson, 1993). Scales for constructs of interest in the workplace such as integrity tests can be composed of items that assess compound personality traits. For example, 'ambition' can be understood as a compound trait because it is composed of aspects of Conscientiousness and Extraversion (Hough & Ones, 2001). In the workplace 'customer service orientation' has been shown (Ones & Viswesvaran, 2001) to be a compound trait consisting of the Big Five dimensions of Agreeableness, Conscientiousness, and Emotional Stability (Neuroticism reversed scored). 'Managerial potential' is a compound trait arising from Extraversion, Emotional Stability, and Conscientiousness (Hough, Ones, & Viswesvaran, 1998). The criterion related validities associated with broad, compound personality variables are substantial, higher than those reported for any of the bivariate correlations of any one of Big Five with the criterion of overall job performance (Ones et al., 2005).

Another area of relevance to work and organisational settings in which the study of personality has been fruitful is in the area of career progression and success (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007). There are two dimensions to this aspect of an individual employee's life, namely, job satisfaction (intrinsic), and aspects such as salary and position in the organizational hierarchy (extrinsic). Four of the Big Five dimensions have been shown to relate to either extrinsic or intrinsic career success, with Conscientiousness and Extraversion being associated with slightly higher levels of extrinsic and intrinsic career success, and Neuroticism and Agreeableness being associated with slightly lower levels of career success (Judge,

Higgins, Thoresen, & Barrick, 1999). These insights are of use when it comes to career counselling of employees. The effect sizes are small because of the role that moderators such as family status or industry characteristics play in determining career outcomes, and there are many contingencies that might alter the relationship between personality and career outcomes.

The foregoing is a brief overview of the importance of the dimensions of the Big Five in understanding human behaviour in the workplace and in organisational settings. From a construct validity perspective the accurate measurement of these dimensions is critical when it comes to criterion related validity and the assessment of each individual employee's standing on these dimensions (Salgado, 2016). The research underlying the findings described above is generally based on participants who are job incumbents rather than job applicants in many studies of the role of personality. For organisations to benefit fully from advances in both psychometrics and the understanding of construct validity, as described in Chapter 2, it is of vital importance that the assessment of personality for selection purposes be accurate. Otherwise it will not be possible for organisations to fully benefit from the insights available from the accumulated body of research on personality in work and organisational settings. Having examined some of the evidence for the importance of personality in the workplace, the next sections review research on the hierarchical structure of the Big Five personality dimensions in the following sections.

3.3 The Higher Order Structure of Personality

There has been an on-going debate within the field of personality psychology regarding the hierarchical structure of personality superordinate to the five broad

dimensions. An answer to this question has important methodological implications for the hypotheses tested in this research programme. A ‘strong’ approach to construct validity is taken in this chapter in order to meet the Messick’s (1995) criteria for establishing construct validity of the NEO PI-R in the context of high stakes employee selection contexts. According to Strauss and Smith (2009), “Strong programs depend on precise theory, and are perhaps accurately understood to represent an ideal. Weak programs, on the other hand, stem from weak, or less fully articulated, theories and construct definitions” (p. 9).

Even though the Big Five dimensions were originally conceptualised as orthogonal constructs, with little to no shared variance across the five factors, there is now a body of empirical evidence from factor analysis that has been used to argue for a smaller number of higher level factors, or metatraits, organised hierarchically which putatively explain the variance that the five lower order dimensions of personality have in common (Chang, Connelly, & Geeza, 2012). These factors are referred to as Alpha and Beta (Digman, 1997), or Stability and Plasticity (DeYoung, 2006) by different researchers. In factor analytic studies Neuroticism, Agreeableness, and Openness load on both Alpha and Stability. Extraversion and Openness load on both Beta and Plasticity. They represent the manifestation in personality of the two broadest requirements of any human being.

There is also some support for the existence of a putative single higher second order factor at the apex of the hierarchy of personality dimension– the general personality factor GFP (Just, 2011; Loehlin & Martin, 2011) – which is superordinate to Stability and Plasticity, accounting for the common variance in Stability and Plasticity. Alternatively, some researchers maintain that the GFP itself exists as a first order higher factor which better accounts for the common variance of the five broad

dimensions than the putative Stability and Plasticity factors (Musek, 2007; van der Linden, Nijenhuis & Bakker, 2010). Support for the existence of a GFP comes from a number of more recent empirical studies (Musek, 2007; Ruston & Irwing, 2008; van der Linden, Bakker, & Serlie, 2011).

However, a number of researchers hold the view that these higher order factors are no more than methodological artefacts due to factors such as socially desirable responding that arise as a consequence of assessing personality using self-report measures. This could also be due to other non trait variance due to contaminants such as the ‘evaluative’ content of items in the various personality inventories in use (Anusic, Schimmack, Pinkus, & Lockwood 2009; Ashton et al., 2004; Hopwood & Donnellan, 2010). Resolution of the debate concerning the higher order structure, if any, superordinate to the Big Five plays an important part in the methodological approach in this research programme. For this reason the following subsections examine the putative higher level structures that have been proposed by different researchers for the structure of personality superordinate to the Big Five.

3.3.1 What are the Higher Order Factors?

According to Digman (1997), the higher order factor Alpha can be viewed as a broad collection of traits that are socially desirable. Hostility, neuroticism, and heedlessness are undesirable traits in any society, whereas Agreeableness, Emotional Stability, and Conscientiousness have long been the subject of moral lessons. Alpha represents the socialisation process itself i.e. the development of impulse restraint and conscience, and the reduction of hostility, aggression, and neurotic behaviour. Similarly, DeYoung’s (2006) Stability reflects an individual’s ability and tendency to

maintain stability, and to avoid disruption in emotional, social, and motivational domains (DeYoung, et al., 2002; DeYoung, 2006). Digman (1997) saw Beta as being related to Maslow's self actualisation concept and Carl Rogers's concept of personal growth encompassing an enlargement of the self by a venturesome encounter with life and its attendant risks, by being open to all experience, especially new experience, and by the unfettered use of one's intelligence. According to DeYoung (2006), Plasticity reflects the ability and tendency to explore and engage flexibly with novelty, in both behaviour and cognition. DeYoung's (2006) descriptors of Stability and Plasticity are used for the rest of this thesis rather than Digman's (1990) original terminology of Alpha and Beta simply because his descriptors convey more meaning than those of Digman.

Rushton described the person high in GFP as "altruistic, emotionally stable, agreeable, conscientious, extraverted, intellectually open, and mentally tough with high level of well-being, satisfaction with life, self-esteem, and emotionally intelligent," (p. 473). In contrast, a person low in GFP is generally maladjusted and likely to have a personality disorder (Templer, 2013). According to Musek (2007), the GFP may reflect a psychobiological disposition that produces the relevant covariations in affective-motivational bases of personality and consequently influences the emotionality, motivation, well-being, satisfaction with life, and self-esteem.

The different putative higher order structures of the Big Five that have been investigated are illustrated below in Figures 2, 3, and 4. As mentioned above one putative explanation for the common variance, or communality, that Stability and Plasticity might share would be the existence of the common factor GFP superordinate to them (Chang, Connelly, & Geeza, 2012). Alternatively, the shared

variance of the Big Five could be explained either by two higher order factors, Stability and Plasticity, or a single higher order factor GFP (Chang et al.).

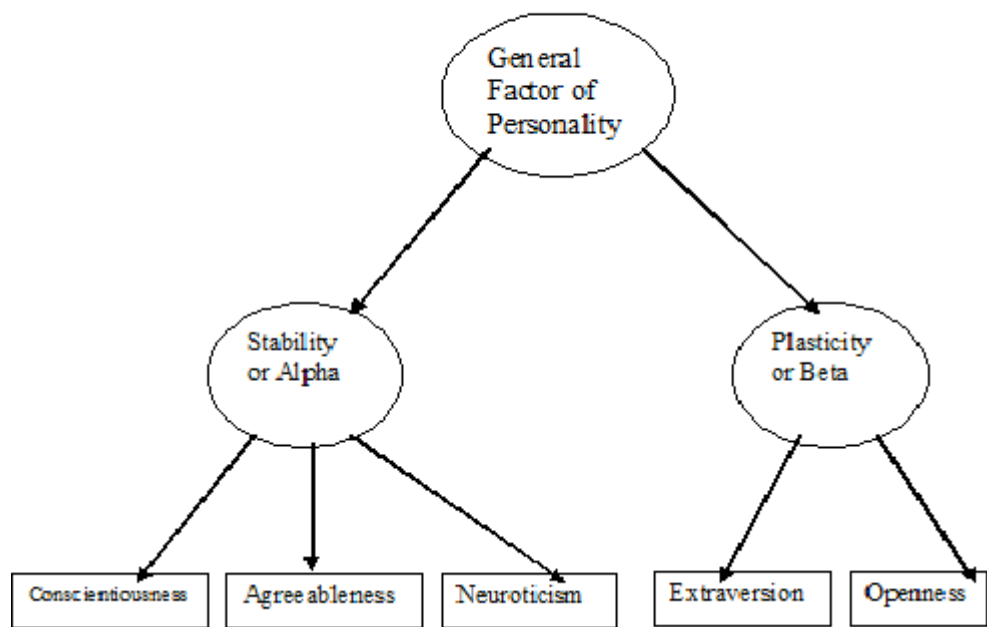


Figure 2 *Big Five Higher Order Putative Structure, with First Order Factors Stability and Plasticity loading on a General Factor of Personality*

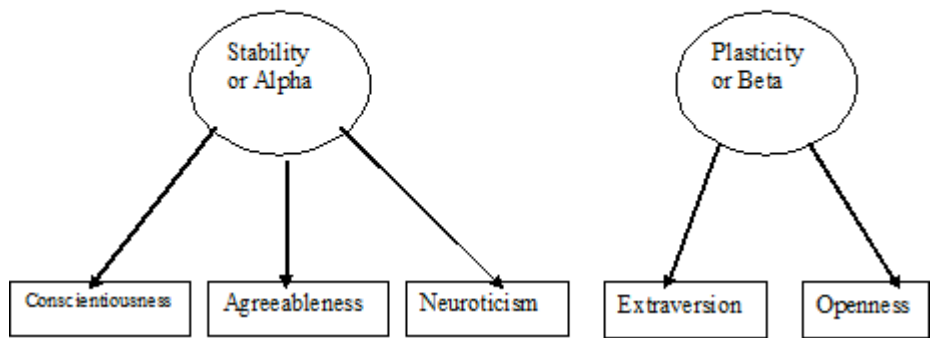


Figure 3 *Big Five Higher Order Putative Structure, with Stability and Plasticity and no Second Order General Factor of Personality*

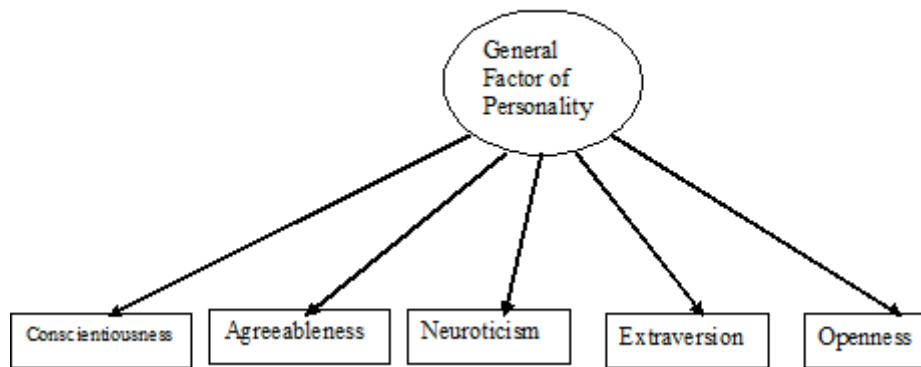


Figure 4 *Big Five Higher Order Putative Structure, with the Big Five Factors loading on a First Order General Factor of Personality*

The difference between these putative higher order structures is that, in the first case, there are two levels of higher order factors, and that the existence of a GFP as well as Stability and Plasticity is not mutually exclusive. This is because the GFP exists at a higher level superordinate to the first order hierarchical factors of Stability and Plasticity (Brown, 2006). In the case of the second putative higher order structure the GFP and the Stability/Plasticity higher order structures are mutually exclusive as shown in Figures 3 and 4. Of course if the shared variance of the Big Five is entirely due to variance arising from common method variance (CMV) then there can be no meaningful i.e. construct valid, higher order factors superordinate to the Big Five dimensions.

3.3.2 The Unbalanced Nature of the Higher Order Structure

The higher order structure of the Big Five was investigated by Markon, Kruger and Watson (2005) in order to examine whether there was factor invariance between normal and abnormal personality populations as part the DSM 5 revision project undertaken by the American Psychiatric Association. They found that the hierarchical structure superordinate to the Big Five was ‘unbalanced’. A balanced hierarchy can be defined as one in which every object at a given level of the hierarchy is at the same level of abstraction, according to Markon et al. (2005). In contrast, an unbalanced hierarchy is one in which objects at a given level of the hierarchy differ in their level of abstraction. According to the authors, “The unbalanced nature of personality hierarchy potentially has methodological implications as well, including implications for psychometric analysis and measure construction” (p. 150). The unbalanced nature that they found can be seen in Figure 5 where Neuroticism (Negative Emotionality) appears at two higher order levels, and there is an additional level between the Big Five and the higher order factors of Stability and Plasticity. In their meta-analysis, Markon et al. (2005) attempted to delineate a hierarchy that would account for variation across the domains of normal and abnormal personality. They found that the Big Five model of personality is only partially isomorphic with their structural model of abnormal personality as shown in Figure 5. Nonetheless each of the Big Five dimensions provides information about normal as well as abnormal personality traits, suggesting that the five-factor level represents an important focus for research on psychopathology and personality, and constitutes a set of ‘building blocks’ for superordinate personality structure.

The importance of the Markon et al. (2005) research is that it sheds additional light on the complexity of the nomological net of the Big Five. This unbalanced structure has methodological implications for using factor analysis when modelling the hierarchical structure of personality. Guastello (1993) described this situation thus, “The architecture of the Big Five is lopsided: The ‘third floor’ does not extend to all wings of the castle. There is no mathematical reason why a natural structure should be constructed in a whole number of dimensions (floors of a building, to continue with the metaphor for vertical structure” (p. 1299).

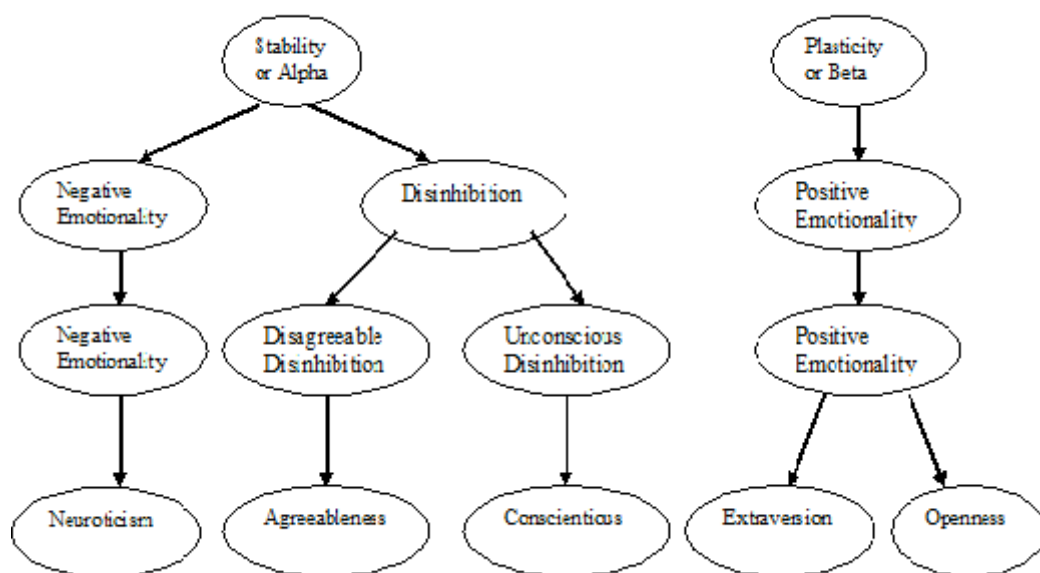


Figure 5 *Markon et al.'s (2005) Unbalanced Big Five Higher Order Structure*

Markon et al.'s research can help to inform the inferences drawn in research from a methodological perspective. It may be of importance when it comes to dealing with the putative presence of correlated errors in any confirmatory factor analysis or

structural equation modelling of the putative higher order structure of the Big Five, and evaluating factor loadings. To summarise the contents of Section 3.3 the questions surrounding the existence of these putative higher order factor models of personality demonstrate that one of the essential requirements for establishing the construct validity of the factor structure of personality (Loevinger, 1957; Messick, 1995) has not yet been fully resolved. The research programme of this thesis clarifies the issue and in doing so contributes to a resolution of the debate.

3.4 The Lower Order Structure of the Big Five

There is a view that each of the Big Five dimensions have two ‘aspects’ of each of the Big Five that are to be found at an intermediate level between the facets and the five broad dimensions (Möttus, Kandler, Bleidorn, Riemann, McCrae, 2007; Sun, Kaufman, & Smillie, *in press*; van Doorn & Lang, 2010). The Big Five Aspect Scale (BFAS) is an example of this - it is used to measure two ‘aspects’ from each Big Five domain, which are distinct from each other both conceptually and empirically (DeYoung, Quilty & Peterson, 2007). The research findings of DeYoung et al. (2007) indicate the existence of two ‘distinct (but correlated) aspects’ within each of the Big Five dimensions, representing an intermediate level of personality structure between facets and domains. For example, the Big Five factor of Extraversion was shown by DeYoung (2006) to consist of the two correlated aspects, which are termed Enthusiasm and Assertiveness. Neuroticism was shown to consist of two correlated aspects, called Volatility and Withdrawal. More recently a number of researchers have made a case for nuances of the facets as single items or groups of single items in a facet scale (Möttus, Kandler, Bleidorn, Riemann, McCrae, 2017).

In particular, support for the existence of the two aspects lower order of the dimension of Extraversion comes from the so-called psychobiological model of Extraversion (Depue & Collins, 1999). According to Depue and Collins (1999), Extraversion can be shown to have two central characteristics. The first one is *interpersonal engagement*, which consists of affiliation (enjoying and valuing close interpersonal bonds, being warm and affectionate) and agency (being socially dominant, enjoying leadership roles, being assertive, being exhibitionistic, and having a sense of potency in accomplishing goals). The second is *impulsivity*, which emerges from the interaction of extraversion and a second, independent, trait which Depue and Collins (1997) refer to as constraint. According to them, agency is a more general motivational disposition that includes dominance, ambition, mastery, efficacy, and achievement. This research is important because, for example, the omnibus NEO PI-R and the Hogan Personality Inventory (HPI) differ in how Extraversion is assessed. The HPI personality measure treats Extraversion as consisting of two dimensions - Sociability and Ambition - each with a number of facets, whereas the NEO PI-R assesses Extraversion as a single broad Big Five dimension consisting of six facets.

As can be seen from this Section and Section 3.3 there are several issues regarding the nomological net of the Big Five dimensions of personality. The importance of these issues, with respect to establishing construct validity of the NEO PI-R in high stakes employee selection situations, will be seen later when it comes to the Analysis section of Chapter 8 of this thesis. Without a proper understanding of such properties as 1) the complexity of factor loadings of items in a personality measure, 2) the existence or otherwise of both higher order factors and lower order aspects of the Big Five, and 3) the unbalanced nature of the higher order structure it

would not be possible to adequately interpret the existing theoretical and empirical ‘fuzziness’ surrounding the nomological net of the Big Five.

3.5 The Evidence for a Hierarchical Structure

Following Digman’s (1997) publication of his research, which showed that putatively there were two higher order factors superordinate to the Big Five based on the pattern of correlations reported in 14 studies employing various Big Five instruments and both self and observer ratings, DeYoung et al. (2002) replicated Digman’s two factor higher order hierarchical model of personality. There is also good meta-analytic evidence that the higher order factors of Stability and Plasticity, which are largely uncorrelated, do exist (Chang, Connelly, & Geeza, 2012), as well as multitrait-multimethod (MTMM) evidence (DeYoung, 2006). The findings of Chang et al. (2012) have been supported by the more recent research of Gnambs (2015) whose MTMM studies showed that there was very little evidence for a GFP when the length of acquaintance was taken into account.

The publication of Musek’s (2007) paper was the first to make the case for the existence of a GFP which in turn has led to a growing body of research into the putative higher order structure of personality with an emphasis on the evidence both for and against the existence of a GFP (Chang, et al., 2012; Hopwood & Donnellan, 2010; Just, 2011; Rushton et al., 2009; Schermer & Vernon, 2011; van der Linden, Nijenhuis & Bakker, 2010).

.Key to understanding the disagreements between advocates of the putative existence of a GFP and those who disagree is the finding of correlated latent factors.

This points to shared variance between the latent factors and could be suggestive of another factor one level higher in the hierarchy (Brown, 2006). There is one important caveat to this and that is that if the shared variance is due to measurement error, or some other artefact such as CMV, then the higher order factor ‘discovered’ is not a substantive factor but rather a method factor due to statistical/methodological artefacts (Chang, Connelly, & Geeza, 2012).

3.5.1 Review of the Research in support of a GFP

A paper published by Musek (2007) was the first to draw attention to the possible existence of a GFP. This question has relevance to the evaluation of the hypotheses tested in the research programme. The methodology used by Musek (2007) is of interest because his research has played a pivotal role in the search for evidence for a GFP. The data for his study were collected using three different samples, with Sample 1 (N=301 adults) completing the Slovenian versions of the Big Five Inventory (John, Donahue, & Kentle, 1991), Sample 2 (N = 185 adults) completing the Slovenian version of Goldberg IPIP 300 Items Questionnaire, and Sample 3 (N=285 adolescents) completing the Slovenian version of the Big Five Observer (Caprara, Barbaranelli, & Borgogni, 1994). It is important to note that each of the three samples studied were monomethod studies with no procedural or statistical precautions taken to deal with, and/or detect CMV. Musek relied on confirmatory factor analysis (CFA) to provide what he referred to as ‘decisive evidence’ for the existence of a GFP. The primary CFA analyses carried out showed poor fit even when judged by the more lenient approach to judging fit recommended by Hopwood and Donnellan (2010).

The only models tested by Musek that provided acceptable fit resulted when post hoc a number of modifications, suggested by the Modification Indices provided by the CFA software (Brown, 2006; Byrne, 2010; Kline, 2011), were included in the models tested. This is not unreasonable since at the most basic level it is very difficult to write 'perfect' items for assessing personality that only load on one primary factor with no secondary factor loadings. This means that some items may unavoidably tap additional if substantially minor sources of variation (Johnson, 1995). These sources of variance frequently can lead to correlated residuals in a CFA analysis and affect the overall model goodness of fit when not explicitly included in the analysis (Hopwood & Donnellan, 2010). The theoretical reason for applying these modifications to the confirmatory model put forward by Musek (2007) was that "they were justified on the basis of assumed covariations between the errors of variances produced by the influences of social desirability and semantic similarity" (p. 1223). He then went on to state that "after the reduction of degrees of freedom to 3, further modifications brought no substantial increase of fit indices". This suggests that the modifications to the initial confirmatory model were determined more by the modification indices than any theoretical justification.

The use of modification indices to improve model fit must be treated with caution. Strictly speaking it is only when they are justified based on prior theory that they should be allowed (Brown, 2006, Kline, 2011). Each of the final models in three samples evaluated included two correlated errors - between Extraversion and Openness in all three models, between Neuroticism and Openness for BFI data, between Conscientiousness and Agreeableness for IPIP-300 data, and between Conscientiousness and Neuroticism for BFO data (Musek, 2007). The final model tested for one of the two adult samples had factor loadings for Agreeableness and

Openness that were negligible. These results, together with the post hoc modifications to the original models tested, raise doubts about the findings of Musek's (2007) studies. Brown (2006) makes the important point that "accordingly, re-specified models should be interpreted with caution. Especially in instances where substantial changes have been made to the initial model" (p. 124).

In his discussion Musek (2007) strongly asserts that "According to the results obtained from the confirmatory analyses, the presence of one common and general highest factor in the Big Five personality space is beyond doubt" (p. 1225). This assertion has been rejected by Comensoli and MacCann (2013) and by Lance and Jackson (2015) on methodological grounds. Therefore the assertion by Musek is, arguably, an unsustainable assertion when the CFA findings are taken into account, and when these CFA results are critically examined from the perspective of the approach to construct validity of, for example, that of Messick (1995) and Embretson (2007). Musek (2007) did not perform a higher order CFA on the three samples he used. Using the data – correlation matrix, standard deviations, and means - contained in the Musek (2007) article higher order CFA's were carried out, as part of this research programme, using AMOS 23 on this data. The poor results obtained are consistent with the questions raised by Comensoli, & MacCann (2013) concerning his finding that the existence of a GFP was 'beyond doubt'. Finally, Musek (2007) relied on the finding of an EFA dominant general factor to make the case for the existence of a GFP. However, as Brown makes clear, "A limitation of EFA is that its identification restrictions preclude the specification of correlated errors. Thus, the results of EFA may suggest additional factors when in fact the relationship among some indicators are better explained by correlated errors from *method effects*" (p. 159). In addition, Lance and Jackson (2015) make the point that a dominant general

factor “is guaranteed to emerge from an orthogonal PCA decomposition of *any* correlation matrix” (p. 453). This criticism of Musek’s (2007) research findings raises the issue of the structural fidelity aspect of construct validity (Loevinger, 1957; Messick, 1995). The results of the CFA’s of the three Musek (2007) samples, each with a different Big Five measure, clearly showed that in the sample using the long form Big Five the internal structure measure used differed greatly from that of the samples using the two short forms samples. This then raises questions about the generalisability (Messick, 1995) of his findings to different populations, and the meaning (Embretson, 2007) of the internal structure of the three measures used in his research.

In spite of this it has been argued that data from a number of thirteen published papers containing diverse measures of personality provide very strong support for the position of the GFP at the apex of the hierarchy of personality structure (Just, 2011). Comensoli and MacCann (2013), on the other hand, are very critical of the conclusions arrived at by Just (2011) in her review of the published papers from a number of researchers in support of a substantive GFP. They make the case that when the articles that Just (2011) relies on to support her contention that there is a GFP are closely examined a number of psychometric inadequacies are to be found. These include the overuse of model modifications in the structural equation models, (b) inconsistent application of statistical procedures, (c) the use of secondary data that limit opportunity for cross-validation, and (d) limited exploration and reporting of theoretically important models. In one of the studies reviewed by Just (2011) she referred to a Rushton, Bons, Ando, Hur, Irwing, Vernon, and Barbaranelli’s (2009) article which examined the evidence for a GFP derived from sixteen data sets which used four different personality measures. However, the

number of studies referred to is misleading in that only six of the data sets dealing with the one of the personality measures were from independent samples, because many of the data sets included were based on the reuse of the same sample (Comensoli & MacCann, 2013). In addition, just as with the Musek (2007) study, many post-hoc modifications were required in order to achieve adequate fit of the intermediate solutions that were reported. For example, the CFA analysis of the one of the personality measures examined by Rushton and Irwing (2009) involved fixing three factor loadings and adding two correlated errors in order to obtain adequate fit. This methodological approach raises doubts about the validity of the conclusion reached (Brown, 2006; Byrne, 2010; Kline, 2011).

Most of the studies that support the existence of a GFP have evaluated one personality measure in isolation. Hopwood, Wright, and Donnellan (2011) looked at eight different personality measures in order to see if there was a single GFP common to the eight different measures. They used three different analyses to try to see if there was a common GFP – a principal axis EFA of the factor scores of the personality measures, a joint EFA of the facet scales of three of the omnibus personality measures, and a CFA analysis of each personality measure. If there is a GFP then, it was argued, there should be a higher-order factor from different broad-band personality inventories that should show convergent validity across measures. Based on this Hopwood et al. (2011) argued that there should be strong convergence across the eight different personality measures they looked at, if there was a GFP. They failed to find convergence across these measures. The fact that Hopwood et al. (2011) failed to find support for a GFP that converged across instruments casts doubt on its importance as a substantive personality construct. They state that “Overall, the failure to find convergence across these measures in this study casts significant doubt on the

importance of the GFP” (p.477). This is an important finding in that it challenges the position of Rushton (2012), who listed 24 different personality inventories which had been investigated and found by him to support the existence of a GFP. The Hopwood et al. (2011) findings are consistent with those of Zawadzki and Strelau (2010), who analysed 32 facets from the Big Five domains culled from six different personality inventories and also found no evidence for a GFP. In a comprehensive technical analysis of a number of personality measures Revelle and Witt (2013) showed that the existence of a GFP was ‘much muddier in personality’ (p. 503) compared with general cognitive ability. They recommended that research efforts be directed to developing theory at the lower order level. Their findings have not been challenged.

In summary, this section showed that there have been some serious questions raised by researchers concerning the putative existence of a GFP. The questions raised are methodological in nature rather than theory driven. The importance for this research programme of establishing the existence or otherwise of a GFP is covered in the next sub section.

3.5.2 Relevance to the Research programme

The aim of the present research is to investigate the construct validity of the NEO PI-R in ‘high stakes’ employee selection situations. The research was undertaken in order to examine different aspects of the substantive, structural, and generalisability components of construct validity by investigating the existence or otherwise of the putative higher order factors superordinate to the Big Five, as measured by the NEO PI-R, using a sample of subjects from a high stakes occupational setting.

High stakes employee selection is normally a monomethod assessment situation, at the applied level, in which there is a potential reward to be gained by job applicants. If applicants respond to the items in the personality inventory by using a strategy of faking good then they may improve the odds of being selected.

It should be possible to more reliably evaluate, in a monomethod study, the putative existence of the higher order factors of the Big Five by the use of procedures designed to minimise the effect of CMV due to socially desirable responding (Podsakoff, MacKenzie, & Podsakoff, 2012). Arising from this possibility, and by making use of the ability of Campbell and Fiske's (1959) MTMM approach to separate substantive trait variance from method variance, a comparison of the higher order structure of personality obtained for the NEO PI-R from the participants in the field study of this research should be possible using the findings about the higher order structure from a number of extant MTMM studies of the structure of personality that have recently been carried out. This methodological approach can therefore be used to test the hypothesis that the use of a formal verbal warning about procedures to detect impression management, prior to completing the NEO PI-R, is effective in minimising or eliminating impression management by job candidates resulting from socially desirable responding in the form of faking good in a monomethod study. If the faking good warning is effective then contamination of the substantive Big Five latent constructs by CMV due to this form of socially desirable responding could have been eliminated. The results of the monomethod field study can then be usefully evaluated by comparing them with the findings of the extant MTMM studies of the higher order structure of personality. This is the first study to take this methodological approach to evaluating the effectiveness of a formal warning in personality assessment in high stakes selection.

Using this methodological approach, the argument is made, based on McDonald (1999) discussion of MTMM matrices, that it will be possible to provide a methodologically sound comparison with which to evaluate the construct validity of the NEO PI-R, when used to make inferences about candidates in a high stakes employee selection situation. The generalisability aspect of Messick's (1995) approach to validity examines the extent to which score properties and interpretations generalise to and across population groups, settings, and tasks. A failure to show that the findings concerning the higher order factor structure arising from MTMM studies generalised to the field monomethod study, that is the subject of the research undertaken in this thesis, would call into question inferences made about the construct validity of the NEO PI-R when used in monomethod high stakes employee selection situations even with the use of a pre-assessment formal verbal warning.

It is worthwhile recalling that Messick (1989, 1995) described the substantive component of construct validity as referring to the theoretical rationales for the observed consistencies in test responses, along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks. If a large portion of the variance can be attributed to CMV due to socially desirable responding by participants in the field study, then the substantive component of construct validity of the NEO PI-R when used in high stakes employee selection contexts is obviously compromised. In addition, at the aggregate level the structural component may be compromised due to the fact that the factor structure obtained in the field study may differ from the factor structure of the construct domain of the NEO PI-R in other settings and contexts (Schmit & Ryan, 1993). Even if the factor structure turns out to be invariant the consequential aspect of construct validity may still be compromised due to biased and unfair rank order effects (Hollenbeck, 2007).

It is important to note here that the NEO PI-R is widely used in occupational settings such as career counselling as well as clinical settings (Costa & McCrae, 1992). The effect of faking good on the construct validity of the NEO PI-R in these contexts is not being investigated, or questioned, in this research programme.

The review of this section shows that there is strong support for a higher order structure superordinate to the Big Five consisting of the two latent factors Stability and Plasticity compared with methodologically questionable evidence for a GFP. Consequently, a comparison of the higher order factor structure of the Big Five from the monomethod field study, of this research programme, with extant MTMM research will be used to determine if a formal warning was effective in at least minimising faking good. This is because extant MTMM research separates CMV from substantive trait variance.

3.6 Additional Measurement Issues

The previous section discussed the issues of the higher order structure of personality, and how knowledge of this is important in evaluating the construct validity of the NEO PI-R. This is important from an applied perspective. The NEO PI-R it is probably the most popular of the personality inventories based on the Big Five, and is widely used in occupational contexts (Salgado, 2016). Faking good may represent a significant threat to the construct validity of the NEO PI-R when it is used in applied occupational settings (Caldwell-Andrews, Baer, & Berry, 2000).

In addition to this, there are some other issues that are relevant from a construct validity perspective including the relationship between reliability and

number of items in a measure (McDonald, 1999), and the effect of crossloading of items on latent factors in CFA (Brown, 2006, Kline, 2011). There are, also, other omnibus personality inventories based on the Big Five available in use today which measure the five broad dimensions. These include the Hogan Personality Inventory (HPI) and the Personal Characteristics Inventory (PCI), among others. As previously discussed, each widely used measure of the Big Five is somewhat idiosyncratic in that each has different facets that make up the particular personality measure's summed score for each of the Big Five dimensions (Hopwood & Donnellan, 2010). These inventories have been shown to be congruent but to some extent they each measure their own Big Five rather than THE Big Five, in that they are not parallel measures much less tau equivalent or strictly parallel (Hopwood, Wright, & Donnellan, 2011). They are best described as alternate form measures of personality, and care must be exercised when comparing the assessment results using different measures of the Big Five..

It should be noted, too, that there are also a number of 'short' self-report questionnaire measures of the five factors such as Costa and McCrae's short version of the NEO, namely, the NEO-FFI with 60 items (Costa & McCrae, 1992) and John's Big Five Inventory (BFI) with 44 items (John, Donahue, & Kentle, 1991), both of which measure the five broad dimensions without assessing any subordinate facets (John & Srivastava, 2008). By comparison the NEO PI-R has 240 items. Empirical research has been indeed conducted on the higher order structure of the Big Five using measures containing as little as one item per facet to eight items to measure each dimension (Hopwood & Donnellan, 2010), even though the Spearman Browne 'prophecy' formula clearly shows that when every item in a personality measure has the same true and error variance the reliability of a test, of a certain number of items,

is a simple increasing function of the number of items (McDonald, 1999; p. 94). The issue of reliability of a measure is a very important one when it comes to using factor analysis.

Yet another issue that has to be considered when it comes to factor analysis is that of cross loading indicators (Kline, 2012). Ideally, in constructing a psychometrically sound test the items should be homogenous in that strictly speaking, the measure of each construct of interest should be unidimensional rather than multidimensional i.e. from a factor analytical perspective each item should load only on the latent factor of interest (Clarke & Watson, 1995; McDonald, 1999). This ideal can be difficult to achieve at times, particularly so in the case of omnibus personality inventories with a number of factors. For instance, the AB5C model of personality is based on research which shows that most of the items in personality questionnaires have loadings on more than one of the five factors (Hofstee, De Raad, & Goldberg, 1992). Many items have a secondary loading on another factor as well as a primary loading on the main factor of interest (Johnson & Ostendorf, 1993) – they make the point that “We have shown, however, that scales proposed by different researchers to assess the Big Five often tend to be blends rather than pure markers of these factors” (p. 574). Johnson (1994) looked at the relationship between two omnibus personality inventories - Costa and McCrae’s (1992) NEO PI-R, Hogan’s (Hogan & Hogan, 1995) HPI, and the AB5C inventory (Johnson & Ostendorf, 1993) - and confirmed this property. In AB5C terms, a trait’s facets are depicted by the two factors, a primary and a secondary, that best describe it. Johnson (1994) showed, for example, that with respect to Agreeableness in the NEO, and Likeability in the HPI, that the primary and secondary factor loadings for the NEO PI-R and the HPI were:

Table 1
Comparison of NEO PI-R and HPI facet loadings

NEO PI-R (Agreeableness)		HPI (Likeability)	
Trust	A+N+	Easy –to-live-with	A+E+
Straightforwardness	A+C+	Sensitive	A+O+
Altruism	A+E+	Caring	A+E+
Compliance	A+A+	Likes People	E+A+
Modesty	A+N-	No Hostility	A+N+
Tendermindness	A+E+		

The two omnibus inventories that Johnson (1994) examined have different facets that go to make up the five broad personality dimensions, as can be seen above. For each facet of the broad dimension the primary factor loading is signified first and the secondary loading follows that in the notation, as shown in Table 1. The + and – negative sign indicates the pole of the dimension on which the facet loads. For example, the Agreeableness facet of Modesty in the NEO PI-R has a primary loading that is high on A (Agreeableness) as well as a secondary loading that is low on N (Neuroticism). This lack of unidimensionality or homogeneity is a problem from the perspective of both the substantive and structural components of Messick’s (1995) unified construct validity approach when relying on techniques such as confirmatory factor analysis to investigate these components. It also highlights Loevinger’s (1957) ‘massive systematic distortions’ (p. 646) of measurement problem in which measurement errors which are correlated with true scores which are not random. Demonstrations of negligible relationships with known sources of distortion are an *essential* rather than optional step in test validation (Kane, 2013). This is a ubiquitous problem when it comes to determining the internal factor structure of an omnibus personality measure using confirmatory factor analysis (CFA), and will be referred to in detail in Chapter 8 in the evaluation of the CFA models tested.

To summarise the contents of this chapter the review of the literature showed that the structure of personality is usefully described by the broad Big Five dimensions. There are still some questions concerning the putative higher order structure of personality superordinate to the Big Five as shown by the debate among researchers regarding the existence in latent space of two or one higher order latent factors. There is also a debate about the need for an intermediate level of abstraction between the facets and broad dimensions of omnibus measures of personality. The next chapter examines this major issue in some detail.

Chapter 4

Socially Desirable Responding

It was shown in Chapter 3 that there are still some unresolved questions among researchers concerning the structural aspect of measures of personality (Embretson, 1983; Loevinger, 1957; Messick, 1995), which is primarily concerned with the internal validity of a measure. This chapter expands on that issue by reviewing the extant research on the occurrence of socially desirable responding in general, and in self report measures of personality in particular. The topics reviewed in this chapter will help in establishing the construct validity nomological net as defined in Chapter 2 of both the NEO PI-R personality measure and, also, the impression management measure used in this programme of research. Test score interpretations have social consequences (Messick, 1995; Messick, 1998). If the social consequences of the testing context trigger moral hypocrisy in the test taker and the subsequent interpretation of test score ignores this, then construct validity suffers and the inferences made about candidates may well be meaningless (Cronbach & Meehl, 1955; Loevinger, 1957).

The present chapter is structured as follows: Section 4.1 looks in detail at the effect of socially desirable responding such as for example impression management in the form of faking good on the accuracy of measurement of the Big Five. This is followed in Section 4.2 by a detailed review of research findings pertaining to moral hypocrisy from the fields of social psychology and behavioural economics. Subsequently, Section 4.3 reviews research on suggested remedies for dealing with

faking good. Section 4.4 then provides an overview that draws the research from other three sections together.

Following the structure set out in the opening paragraph the rest of the chapter contains a review of the research on the topic with respect to questions concerning both the occurrence of socially desirable responding, and its effect on construct validity in the measurement of the Big Five of personality with particular emphasis on socially desirable responding in the form of faking good, or “situationally specific intentional distortion” (Sackett, 2012; p. 331) in high stakes selection.

4.1 The Effect of Socially Desirable Responding on the Measurement of the Big Five

Socially desirable responding in all its guises can be a major issue in any assessment using self-report measures (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), and especially so in high stakes employee selection situations according to some researchers (Griffith & Converse, 2012). Faking good is a form of socially desirable responding that is both deliberate and situationally specific (Sackett, 2012). According to Berry and Sackett (2009), “Even though personality measures retain much of their validity in the presence of faking, the fact that fakers displace nonfakers is seen as unfair” (p. 837). This highlights the fundamental problem of the ‘weak’ approach of using criterion-related validity indices to establish the construct validity of a measure (Kane, 2013). It ignores the theoretical admonitions of both Cronbach and Meehl (1995) and Loevinger (1957), and the reason why the Campbell and Fiske (1959) multitrait-multimethod (MTMM) approach to separating substantive from method variance is so important. Messick (1995) and Embretson (2007) make specific

recommendations concerning the issue of unfairness arising from a failure to take account of the consequences aspect of construct validity in applied settings aspect.

Questions in personality questionnaires, technically described as ‘items’, have no objectively scored ‘right’ or ‘wrong’ answers, and are usually self-reports of behaviour (McDonald, 1999). This means that, regardless of which of an item’s response alternatives is endorsed by a respondent the test administrator generally cannot be certain if that choice is in fact the ‘correct’ one, or whether some other response is a better description of the person (Paunonen & LaBel, 2012). This is very different to the assessment of cognitive ability where the items can be objectively scored (McDonald, 1999).

Socially desirable responding is typically defined as the tendency to give positive self-descriptions (Tracey, 2016; Ziegler et al., 2012), but it can also manifest itself in clinical settings as a tendency to give negative self-descriptions (Perinelli & Gremigni, 2016); Salgado, 2016; Sollman & Berry, 2011). Paulhus (1984) showed that the person being assessed may either be consciously engaging in a deliberate strategy of misrepresentation to make an impression on those who might eventually see his or her personality profile, or the misrepresentation could occur at an unconscious level and be motivated by a latent need for self-enhancement and ego maintenance. Socially desirable responding can therefore present a problem when accuracy in the assessment of personality is a concern, and from a construct validity perspective it must be taken into account (Dragow, Stark, Chernyshenko, Nye, Hulin, & White, 2012; Ellingson, Heggstad, & Makarius, 2012; Fan, Gao, Carroll, Lopez, Tian, & Meng, 2012; Griffith & Converse, 2012; Holden, 2008; Komar, Brown, Komar, & Robie, 2008; Landers, Sackett, & Tuzinski, 2011; Marcus, 2006; McFarland, 2003; Mueller-Hanson, Heggstad, & Thornton, 2003; Pace & Borman,

2006; Peterson, Griffith, & Converse, 2009; Sackett, 2011; Salgado, 2016). This problem with self-report measures and the link with behaviour was pointed out a long time ago by La Piere (1934) in his classic paper on the relationship between attitude and behaviour. He stated that “Yet it would seem far more worthwhile to make a shrewd guess regarding that which is essential than to accurately measure that which is likely to prove quite irrelevant” (p. 237).

There is disagreement in the I/O literature regarding the nomological net underlying personality assessment in high stakes employee selection situations, (Morgeson, Campion, Dipboye, Hollenbeck, Murphy, & Schmitt, 2007). Mean scores and standard deviations differ in studies comparing job applicants and job incumbents when completing self-report personality measures (Salgado, 2016). Without general acceptance of the same nomological net the construct validity of a measure cannot be established (Cronbach & Meehl, 1955). Common method variance (CMV) was defined by Podsakoff et al. (2003) as the “variance that is attributable to the measurement method rather than to the constructs the measures represent” (p. 879). The question of CMV arising from deliberate socially desirable responding through ‘impression management’, either by ‘faking good’ or even ‘faking bad’, when responding to the items in an omnibus personality inventory has been the subject of much, sometimes heated, debate and research in recent times (Backstrom, Bjorkland, & Larsson, 2009; Dilchert, Ones, Viswesvaran, & Deller, 2006; Morgeson, Campion, Dipoye, Hollenbeck, Murphy, & Schmitt, 2007). It is an important topic when evaluating the construct validity of personality measures, such as the NEO PI-R. The case against the effect of socially desirable responding in high stakes contexts is considered first.

4.1.1 The Case against a Socially Desirable

Responding Effect

With respect to the use of omnibus personality measures in occupational settings, Hogan, Barrett, and Hogan (2007) maintain that the data reported in their study led to the conclusion that “faking on personality measures is not a significant problem in real-world selection settings” (p. 1270). This position is consistent, to some degree, with the view of Ones, Viswesvaran, and Reiss (1996) who, building on their meta-analysis of correlations between personality measures and job performance, concluded that social desirability in the form of faking good does not moderate the criterion-related validities of personality measures in real-world settings after partialing out this form of social desirability from the predictors. They recommended against correcting raw scores for socially desirable responding in the form of faking good in personnel selection situations. An important point to bear in mind is that Ones et al. (1996) were dealing with criterion validity rather than the broader concept of construct validity, described in Chapter 2, with its emphasis on validation being an on-going dynamic and inferential process (Embretson, 2007).

It is arguable that Hogan et al. (2007) did not prove their hypothesis that faking good, in the form of situationally specific intention distortion of responses to items in the personality measure, was not an issue (Landers et al., 2011). Hogan et al. (2007) tried to address criticism of their research through the use of a smaller sample of 141 participants all of whom had completed the personality measure purely for research purposes. Scores for participants who had completed the battery of tests for research purposes were hypothesised not to differ substantially from the scores of those applicants who had completed the personality measure as part of the

employment selection process, and who were retested. They found an average effect size of .057 between the research sample of 141 participants and the retested sample of 5266 participants. Cohen (1992) pointed out that the alpha error, the beta error, power of a test, and effect size are all linked together with a change in one affecting the others. The app G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) allows one to very easily use the approach to evaluating effect sizes recommended by Cohen (1992). G*Power shows that for a two tailed test of independent means and an alpha of .05, power of .9, and an effect size .057 in an equal sample size case they would have needed both samples to be of size 6470 to detect an effect size of .057. For a sample size ratio of 5266/141 i.e. 37/1, they would have not been able to detect an effect size less than .27 in magnitude. Furthermore, in this case they would have needed to have used two samples of size 126218 and 3322 to detect an effect size of .057. It could be concluded from this that Hogan et al. (2007) may well have simply capitalised on chance when they tested their hypothesis using the research sample of 141 participants.

There have been several criticisms of Hogan et al.'s (2007) conclusion. The first was that in their research Hogan et al. (2007) based their conclusion on a sample of job applicants who were not initially selected but were then re-tested six months later. However, the selection decision in the case of the participants was based on the applicants' ability test scores and not their personality test scores, as Landers, Sackett, and Tuzinski (2011) point out. In addition, Hogan et al.'s (2007) conclusion did not build on the re-testing of successful applicants; this is an important caveat with respect to the conclusion they reached (Hausknecht, 2010; Holladay, David, & Johnson, 2013). They may have arrived at a different conclusion if they had retested a sample of successful applicants. This is because those who were successful might

have faked good initially, and on re-testing as incumbents may not have faked good. For example, in their research Donovan, Dwight, and Schneider (2014) found that fakers were disproportionately represented among successful applicants selected compared to nonfakers when the successful applicants were retested later as job incumbents. In this Donovan et al. (2014) study faking by successful job applicants was found to be a common occurrence, with approximately half of the individuals retested being classified as a faker because of the change in their score on the social desirable dimension contained in the self-report measure used.

Holladay et al. (2013) showed that the greatest difference between Time 1 and Time 2 scores in personality assessment of retested job applicants was observed for those who receive failure feedback. Applicants receiving no feedback about the reasons for why they were not selected for the job showed relatively less change in scores across administrations. As has been pointed out the participants in the Hogan et al. (2007) study did not received feedback. The research findings of Fan, Gao, Carroll, Lopez, Tian, and Meng (2012) also contradict the findings of Hogan et al. (2007), and provide support for the argument of Sackett (2012), and Griffith and Converse (2012), that faking is intentional, situationally induced, and changeable, and that faking is indeed a cause for concern.

Faking good does not appear to affect inferences regarding the criterion related validity of personality measures (Birkeland et al, 2006; Hogan, Barrett, & Hogan, 2007; Landers, Sackett, & Tuzinski, 2011; Li & Bagger 2006; Markus, 2006; Ones, Viswesvaran, & Reiss, 1996; Paunonen, & LeBel, 2012) as mentioned earlier. Hollenbeck (2009) maintains that for faking to really affect *aggregated* criterion-related validity, given how low the correlations are to begin with, it would have to cause radical changes in rank orders in order to have an effect on criterion related

validity. This is a very important point. Small changes in rank orderings do, however, affect *individual* selection decisions depending upon where the cut score was set, but this is a different matter to the issue of criterion related validity of personality dimensions in the *aggregate* (Morgeson, et al., 2007)

In summary, the Hogan et al. (2007) conclusion that “faking on personality measures is not a significant problem in real-world selection settings” (p. 1270) has been rejected by most researchers (Salgado, 2016). Whatever descriptor is used to describe the behaviour of impression management in the form of faking good, a case can also be made that the issue of intention to distort responses is central to the behaviour when completion self-report measures. In addition, even though Ones et al.’s (1996) meta-analysis of correlations between personality measures and job performance found that, after partialing out social desirability from the predictors, social desirability did not moderate the validities of personality measures in real-world settings, this finding alone does not meet the criteria of Messick (1995) for establishing construct validity. As Hollenbeck (Morgeson, et al., 2007) suggests faking good would have to cause radical changes in rank orders to have an effect on criterion validity because the correlations between predictor and criterion are low. Both this rank order effect and criterion related validity effect are aspects of construct validity as defined by Messick (1995) and Embretson (2007). Loevinger (1957) points out that reliance on criterion-related validity is an ‘ad hoc’ approach to the establishment of validity, “whereas construct validity is the whole of validity from a scientific point of view” (p. 636). This brings us to a review of the evidence that socially desirable responding in high stakes situations should be a concern.

4.1.1.1 The Case for Socially Desirable Responding in the form of Faking Good

Socially desirable responding is more likely to occur in situations where there is a desirable outcome at stake situations such as ‘high stakes’ recruitment (Griffith, Chmielowski, & Yoshita 2007) that can affect the rank order of individual job candidates (Morgeson et al., 2007). This was one of the reasons that Campbell and Fiske (1959) advocated the use of their MTMM approach. They were concerned with “the adequacy of tests as measures of constructs, rather than the adequacy of a construct as determined by the confirmation of theoretically predicted associations with measures of other constructs” (p. 100).

Socially desirable responding in the form of faking good should therefore be a major concern with regard to the measurement of true scores in personality psychology (Backstrom, et al. 2009; Bangerter, Roulin, & König, 2012; Chan, 2009; Morgeson et al., 2007). For example, the transparency of the item content in an omnibus personality measure is one that can lead to an occurrence of this problem. Research has clearly shown that job applicants can make themselves look better on such items if they choose to do so by faking good’ (Dilchert & Ones, 2012). Because of item transparency test takers can develop reasonably accurate hypotheses about what trait an item is tapping into (Schmit & Ryan, 1993). For the purposes of this thesis the following definition of faking, as a form of socially desirable responding, is used (Ziegler, 2011; Ziegler, MacCann, & Roberts, 2011):

Faking represents a response set aimed at providing a portrayal of the self that a person to achieve personal goals. Faking occurs when this response set is

activated by situational demands and person characteristics to produce systematic differences in test scores that are not due to the attribute of interest (p. 8).

The variance in responding arising from socially desirable responding is associated with situationally specific impression management is also described by Sackett (2012) as faking. He identifies a number of factors that a personality test taker's observed score depends on such as a respondent's true trait score, as well as erroneous self-perception and impression management across contexts and within specific contexts. It is beyond dispute that when participants in research studies have been instructed to fake good they can easily elevate their scores on personality measures (Holden & Book, 2012). It has already been pointed out that unlike ability measures which can be objectively scored personality measures are not necessarily objectively scored. The concept of faking on a personality measure has been described as somewhat of a strategic action in nature because the test taker can use the responses to items to portray herself or himself as a certain kind of person on that occasion (Kroger & Wood, 1993).

Socially desirable responding, such as faking and impression management, is also a very important methodological consideration when it comes to factor analysing omnibus personality measures and conclusion drawing about the construct validity of the putative existence of higher order factors (Holden & Book, 2011; Loevinger, 1957). For the purposes of this research programme it is best described as a form of CMV in personality assessment that arises when using self-report measures (Chang, Connelly, & Geeza, 2012). As a consequence, if the covariances between factors are contaminated by the variance due to socially desirable responding then the existence

of meaningful higher order factors super-ordinate to the Big Five can be questioned (DeYoung, 2006). A prime example of this can be seen in the research of Backstrom, Bjorklund, and Larsson (2009) who found support for their hypothesis that “the general hypothesis is that correlation between personality factors at the domain level (the Big Five) can be largely attributed to a general factor, and that this general factor is caused by social desirability concerns activated by the semantic content of the test items” (p. 336). A recent meta-analytic, MTMM, confirmatory factor analysis also raises very serious doubts about the putative existence of a latent general factor because of contamination of the substantive trait variance by method variance (Chang et al., 2012). This issue will be dealt with in detail in Chapter 6.

Method variance with respect to omnibus personality inventories is variance in Big Five scale item responses throughout the measure that is due to the influence of common method bias (Biderman, Nguyen, Cunningham & Nima, 2011). This variance is attributable to the measurement method rather than to the constructs the measures represent (Podsakoff, et al., 2003). According to the results of the research into the issue of CMV and the Big Five by Biderman, Nguyen, Cunningham, and Ghorbani (2011), there is evidence that measurement models of the Big Five should take into account two types of method bias that arise – one general bias factor influencing all items, and a second type of bias factor influencing items worded either positively or negatively (Ashton, Lee, Perugini, Szarota, de Vries, Blas, & De Raad, 2004). However, CMV can also be due to other causes such as acquiescence and nay-saying for example (Costa & McCrae, 1997; Hughes, in press), as well as strategic situationally dependent deliberate faking of item responses (Griffith & Converse, 2012). Method variance in personality assessments therefore can arise from different

sources, all of which arguably can be subsumed under the single descriptor of CMV (Podsakoff et al., 2003) for the purposes of this research programme.

Finally, based on his research, Paulhus (1984) theorised that socially desirable responding had in fact two dimensions. To account for these, he developed a measure of socially desirable responding called the Balanced Inventory of Desirable Responding (BIDR), which assessed the two dimensions – Impression Management (IM) and Social Deception Enhancement (SDE) (Paulhus, 1998; Paulhus, Harms, Bruce, & Lysy, 2003). IM refers to an intentional distortion of self-descriptions in order to be viewed favourably by others. SDE, in contrast, denotes an unconscious propensity to think of oneself in a favourable light. Faking good is deliberate (Griffith & Converse, 2012), and is arguably related to IM unlike SDE, whereas item response distortion arising from SDE is not deliberate (Ellingson, 2012; Lönnqvist, Irlenbusch, & Walkowitz, 2014). Paulhus's (1984) distinction between the two dimensions of socially desirable responding may help to shed light on whether the putative higher order factors, described in Chapter 3, of omnibus personality inventories exist or not. Other measures of socially desirable responding such as the Marlowe Crowne (Crowne & Marlow, 1960) measure and the bespoke 'lie scales' included in some commercial personality measures, which do not distinguish between these two dimensions as the Paulhus BIDR does, arguably do not shed as much light on this issue (Connelly & Chang, 2015). The measurement of faking good as a form of impression management is explored in detail in Chapter 5. For the purposes of this chapter, it is assumed that the validity construct measurement of socially desirable responding in the form of faking good is not subject to question.

4.1.2 To what extent does Faking Good occur?

In a review of the body of research on the prevalence of faking estimated that 30% of job applicants engage in faking behaviour with a range of + or – 10% (Griffith & Converse, 2012). Further evidence in support of this level of faking in psychological assessment that are deemed to be ‘high stakes’ i.e. potential exists for a gain or loss, comes from Hall and Hall (2012) who point out that in the area of neuropsychological assessments where compensation is involved through litigation that a similar level of faking, bad rather than good in this case, is to be found. The figure they give for assessments in criminal cases is 19%. So faking is an issue that should be a major concern and goes to the heart of the construct validity of personality measures and assessments (Griffith & Converse, 2012; Sackett, 2009, Salgado, 2016; Ziegler et al., 2012). The possibility of faking leads to questions concerning personality test scores ‘measuring’ something other than permanent predispositions to behave, such as momentary presentations of self to suit the occasion, which was the concern that Campbell and Fiske (1959) addressed in their seminal paper.

Today there appears to be a dominant consensus view, at least in occupational settings, that:

1. Faking can, and does, occur (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Griffith & Converse, 2012; Landers et al 2011; Markus, 2006, Salgado, 2016)
2. It does not affect the criterion related reliability of personality measures with respect to job performance (Hogan et al. 2007; Li and Bagger 2006; Ones et al 1996), although Salgado (2016) disputes this.

3. It is more likely to occur in ‘high stakes’ situations such as recruitment where there is a desirable outcome at stake - getting a job (Ellingson, 2012; Griffith, Chmielowski, & Yoshita, 2007).

It is important to note that in regard to Point 2 above, the authors are referring to criterion related validity and are not referring to the broader concept of construct validity. Even though a measure can be shown to have criterion related validity this does not mean that the measure has construct validity as defined by Messick (1989, 1995).

Rothstein and Goffin (2006) carried out a comprehensive review of research on the issue of faking in occupational settings that deals with the three points above in some detail. The empirical evidence from this research showed that when subjects are instructed to ‘fake good’ their standing on the personality dimensions is shifted towards the ‘perceived to be more socially desirable pole’ of each of the Big Five dimensions. The re-testing of participants in work situations on omnibus personality measures, such as the NEO PI-R, demonstrates evidence of increases in scores but not to the same extent where subjects are instructed to fake good (Landers et al. 2011). In addition, Rothstein and Goffin (2006) make the important point that faking may be manifested in substantially different response patterns depending on the individual differences of the test takers, and their perceptions of the nature of the job they are applying for. As mentioned earlier, Hollenbeck explains the apparent contradiction between points 1) and 2) above (Morgeson et al., 2007).

Faking studies are either of a between participants or within participants’ format, and there are a number of variations of these basic designs that are used (Burns & Christiansan, 2011; Sackett, 2011; Walmsley & Sackett, 2013). In some

experimental studies, using random assignment, comparisons are made between the scores of participants instructed to deliberately fake good and other participants who are not instructed to fake good. In other experimental studies the same participants are used in both a faking good condition and a no faking condition. In field studies, which are not based on random assignment, there are two different approaches which are used (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006). In some studies job applicants are compared with job incumbents – a between participants approach (Rosse, Stecher, Miller, & Levin, 1998; Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). Alternatively, successful job applicants' scores are compared with their scores on the same tests after they have joined the organization – a within participants approach (Arthur, Glaze, Villado, & Taylor, 2010; Griffith et al., 2007). Another, less common, variation in field studies is to test and at a later time re-test initially unsuccessful job applicants (Hogan et al. 2007). An important point to bear in mind, too, is that in some research studies, where the pressure of socially desirable responding may be weaker than in an applied context, the researcher may not take the effect of social desirability into account when calculating relationships to different criteria.

The laboratory studies clearly show that it is possible, and not difficult, to fake good on personality measures (Barrick & Mount, 1996, Viswesvaran & Ones, 1999). This can be partly due to the overt and clear cut nature of the item content in omnibus personality measures because job applicants can easily make themselves look better on such items if they choose to do so by faking good. The ecological validity of these studies is, however, questionable when it comes to comparisons with field studies where contradictory results have been found (Griffith et al., 2007; Morgeson, et al., 2007; Vasilopoulos, Cucina, & McElreath, 2005). This is an issue concerning the

generalisability of laboratory results to applied selection settings, which is one of the six criteria enumerated by Messick (1995) for establishing the validity of a measure. Arguing that such faking is unimportant, as some do, because it is a constant affecting all dimensions of a personality measure equally is also questionable (Tett & Christianson, 2007).

Estimates from this research as to the extent of the occurrence of faking good vary, as well as the extent to which each of the Big Five is prone to being faked. Hogan, et al., (2007) are of the view that “all faking all the time” (p.1280) which, according to them, means that the issue of faking good is a non-issue. Arthur et al. (2010) showed in their research that all of the Big Five dimensions of personality are prone to faking good. They found that “although most test takers’ scores were stable, fairly sizeable percentages of the test takers displayed evidence of higher Time 1 (high stakes) than Time 2 (low stakes) scores – specifically, 35.81% on Agreeableness, 34.12% on Conscientiousness, 33.11% on Emotional Stability, 35.81% on Extraversion, and 14.53% on Openness” (p. 8). This is in stark contrast to Hogan et al. (2007)’s findings that “Applicants at T1 get the same scores as they do at T2, and both sets of scores are the same as scores for a sample of research applicants” (p. 1280). In their research Griffin et al. (2007) found that “Six out of the ten applicants hired would not have been chosen if their honest scores were used, rather than their faking inflated applicant scores” (p. 350). Donovan, Dwight, and Hurtz (2003) found that for entry level job applicants 50% of applicants indicated that they had exaggerated qualities or characteristics commonly assessed in personality measures.

Further evidence in support of the research which shows that faking good, as measured, can and does occur in high stakes employee selection situations comes

from Higgins's (1987) Self Discrepancy Theory. According to this theory, negative psychological situations can result from discrepancies between an individual's self concept and a significant other's aspirations for an individual. In an employee selection situation, the 'other' is represented by the selecting organisation. According to Higgins (1987), there are three dimensions to the self - the actual, ideal, ought selves – which may not be congruent in the individual thus leading to psychological discomfort. Faking good can ease this emotional discomfort in selection situations. Support for this viewpoint comes from the factor analytic research of Schmit and Ryan (1993) who found that job applicants exhibited a work-related personality factor in addition to the Big Five, which they called an *ideal-employee factor*. The job applicant may respond differently to those items that he or she views as requiring a self-presentation to the significant other that varies from the actual self image of the respondent (Biderman & Nyugen, 2009; Cellar, Miller, Doverspike, & Klawnsky, 1996; Klehe, Kleinmann, Hartstein, Melchers, König, Heslin, & Lievens, 2012). This alternative self-presentation is one that might well involve a self evaluation based on a comparison with competent employees that the applicant knows (ought self) or a comparison to an idealised version of an employee (ideal self). However, Salgado (2016) points out that the ideal-employee factor could also be due to the effect of restricted range sampling.

In summary this subsection reviewed the evidence for, and against, socially desirable responding in the form of faking good. The preponderance of evidence supports the position that it does occur. Its effect on criterion related validity may not be significant. However, it can lead to consequential unfairness and bias when the selection decision is made on a rank ordering of applicants for a position in an organisation which is what occurs in many high stakes employee selection situations.

The next section looks at the research from social psychology and behavioural economics regarding the topic of moral hypocrisy and its possible link with faking good.

4.2 Faking Good - A Form of Moral Hypocrisy

According to Messick (1995), the substantive aspect of construct validity refers to theoretical rationales for the observed consistencies in test responses and empirical evidence that the theoretical processes are actually engaged in by respondents in the assessment tasks. Embretson's (1983) concept of 'construct representations' involves identifying the mechanisms – the processes, strategies, and knowledge - that underlie item responses. This section addresses some of these aspects of the construct validation process. While the consensus in I/O psychology research supports the argument that faking does occur, and that its occurrence should be of concern, this research has tended to ignore research from other areas such as social psychology and behavioural economics as this section of the chapter will show. This research has an important contribution to make and its inclusion helps to add to the nomological net, as advocated by Cronbach and Meehl (1955), underlying personality assessment in high stakes employment situations as well as expanding on both the substantive aspect of construct validation and construct representation.

Faking behaviour is not simply a characteristic of the person who engages in deliberate distortion as a result of a response style (Ayal, Hochman, & Ariely, 2016; Ziegler et al., 2012). As Ellingson (2012) asserts people only engage in faking behaviour when they need to fake. She essentially argues that faking good by job

applicants is simply an example of moral hypocrisy in a particular context. In employee selection situations moral hypocrisy can easily occur as a function of the desirability of the job on offer. Ellingson (2012) enumerates a number of factors, including situational factors, that can determine whether faking occurs or not. It can be a function of the individual's perception of her/his marketability – the greater the degree of freedom with respect to job opportunities that the individual has the less likely he or she will engage in faking behaviour. Job search self efficacy is also a factor – if an individual is confident that the job search efforts will produce positive results he or she is also less likely to engage in faking behaviour. The other important feature of the Ziegler et al. (2012) definition is that it is a goal directed behaviour which has an objective – favourable self presentation in order to achieve a personal goal.

At the most basic definitional level faking on a self-report personality measure is arguably nothing more than a euphemism for lying. Webster Dictionary (<http://www.webster-dictionary.org/definition/Lie>) defines the intransitive verb 'lie' as follows - "to make an untrue statement with intent to deceive" or "to create a false or misleading impression". These definitions are synonymous with the definition of Ziegler et al. (2012). Lying is a form of moral disengagement (Bandura, 2002; Jacobsen, Fosgaard, & Pascual - Ezama, 2017). There is a gap in the extant research concerning faking good in occupational and employee selection situations and the effect of moral hypocrisy on job candidates' responses to items in self-report personality measures. According to Bandura (2002), individuals engage in self regulatory behaviour when they behave in a manner consistent with their adopted standards or right or wrong. The constraints on behaviour are thus the result of each individual's regulation of their behaviour in order to avoid violating their adopted

standards of what is normatively right or wrong. In Bandura's (2002) opinion morality is not rooted in dispassionate abstract reasoning. Instead, it is the result of individuals exercising self influence.

Moral disengagement therefore occurs when individuals deliberately disengage from self censorship, and it involves minimising or distorting the consequences that follow from detrimental actions such as lying. 'Moral justification' and 'euphemistic labelling' are some of the mechanisms of moral disengagement identified by Bandura (2002). Euphemistic language is a key self-deceptive tactic that allows individuals to behave unethically in organisations (Detert, Trevino, & Sweitzer, 2008). These are followed by misconstruing, ignoring, or minimising the consequences of the reprehensible conduct such as euphemistic labelling. To engage in moral disengagement behaviour, such as faking good when responding to items in self-report personality measures, is an example of moral hypocrisy. Not all examples of moral disengagement would be classified as moral hypocrisy. It is arguable that while an event that is morally wrong it involves moral disengagement but it does not involve moral hypocrisy. Moral disengagement explains why some people who generally abide by moral principals of right and wrong are able to engage in unethical behaviour without apparent guilt or self-censure. Moore, Detert, Trevino, Baker, and Mayer (2012) found in a series of experiments that moral disengagement is widespread across a range of organisationally relevant ethical behaviours such as lying, cheating, and self-serving decision making. Faking good on personality measures is arguably an example of moral disengagement and moral hypocrisy co-occurring.

The field of I/O, and occupational, psychology has largely ignored the research on moral hypocrisy and moral disengagement when it comes to discussing

faking in spite of its relevance to the issue of faking good and impression management in personality assessment (Dullaghan, 2013; Ziegler, MacCann, & Roberts, 2012). The recently published book '*New Perspectives on Faking in Personality Assessment*' (Ziegler et al., 2012) contains no reference to the extant body of research on the topic of moral hypocrisy. Yet there is a body of research on this topic, from social psychology and behavioural economics that is extensive and mainly experimental.

4.2.1 Batson's Research on Moral Hypocrisy

In two landmark series of experimental psychology studies Batson and colleagues (Batson, Kobrynowicz, Dinnerstein, Kempf, & Wilson, 1997; Batson Thompson, Seuferling, Whitney, & Strongman, 1999) introduced the concept of moral hypocrisy which they defined as follows: "*moral hypocrisy*: Morality is extolled - even enacted - not with an eye to producing a good and right outcome but in order to appear moral yet still benefit oneself" (p. 1335). The difference between moral disengagement and moral hypocrisy is to be found in the second part of Batson et al.'s (1997) definition. Unlike moral hypocrisy, those who engage in moral disengagement are not necessarily attempting to appear moral while simultaneously engaging in behaviour that is morally lacking.

The experimental paradigm that Batson et al. (1997) used was based on the Dictator Game (Bazerman & Moore, 2012) which presents participants with a moral dilemma played out on a computer. Participants are required to assign themselves and another participant, who is actually fictitious, to either of two tasks that are different (Batson, 2008). Participants are led to believe that the other (fictitious) participant will

not know that they – the actual subject of the experiment - were allowed to assign the task. One task allows the participant the chance to earn a raffle ticket, and has positive consequences in that it is an enjoyable task. The other task has no chance to earn a raffle ticket, and is described in the briefing of participants as dull and boring. Most of the actual participants – 70% to 80% depending on the study – assigned themselves to the more interesting and rewarding task. Yet only 10% of participants believed that assigning the dull and boring task to the other (fictitious) participant was the moral thing to do.

Batson et al. (1997, 1999) selected moral dilemmas that were simple and easy to understand, and where there would be broad consensus about the morally right course of action so that there would be no doubt as to what action would be considered moral in the experiments (Batson, 2002). The experiment was then varied by adding a coin toss to the options that the participants could choose to make the decision as to the allocation of the two tasks. All participants felt that either assigning the positive consequences to the other (fictitious) participant or tossing the coin was the moral thing to do. Yet during the experiment 80% to 90% of participants choose not to flip the coin, and most of them assigned themselves with the interesting and rewarding task. The really interesting finding was that of those who choose to make the decision based on the toss of the coin 85% to 90% assigned themselves the interesting and rewarding task, even though an analysis based on chance says that the figure should have been around 50% for those who tossed the coin. This clearly indicates that moral hypocrisy is occurring with a sizeable number of participants in the experiments. The experiments of Batson et al. (1997, 1999), and others, show that individuals are clearly concerned with serving their own self interest at the expense of others, and many will engage in moral hypocrisy in order to appear to others to be

seen to have behaved in accordance with generally accepted moral standards of behaviour (Batson, 2008).

There are some parallels between Batson et al.'s (1997, 1999) experiments and the situation faced by individual job applicants completing self-report personality measures in that the selection situation is a win/lose situation with some other applicant possibly winning if the first applicant does not fake good or lie. The opportunity for moral hypocrisy is present in faking good situations – the job applicant can deliberately lie which is immoral behaviour regardless of the euphemistic labelling of the lying as faking good or impression management. Hence the coin toss experiment of Batson et al.'s experiments is a very good example of the moral hypocrisy of faking good in that many of those who tossed the coin arguably wanted to appear to have engaged in morally correct decision making while deliberately ignoring the actual outcome of the toss. The percentage figures in Batson et al.'s (1997, 1999) series of experiments for those who selected the coin toss option provide support for the level of faking good that Griffith and Converse (2012) maintain is the norm – 30% of applicants fake good with a margin of error of + or – 10%. Batson et al.'s (1997, 1999) figure for the occurrence of moral hypocrisy is similar to Griffith and Converse's figure when the expected random outcome of 50/50 odds for unbiased coin tossing is taken into account in the Batson et al. experiments.

Batson et al. (1999) also explored the effect of self awareness on the actual behavioural manifestation of moral hypocrisy. They did so by having participants in their high self awareness experiments sit directly in front of a mirror, 60 centimetres away from them. In the low awareness experimental condition, the mirror was turned to the wall. The moral hypocrisy effect was eliminated in the high self awareness condition, but not in the low self awareness condition. As the authors state, "in front

of the mirror the coin became scrupulously fair” (p. 531). The outcome from the coin toss was what was to be expected by chance – half of the task assignments of the ‘other’ participants were to the positive consequences task, and half were to the participants themselves. This is what chance would dictate for unbiased task assignments. This is a very important finding which is relevant to preventing faking good from occurring in high stakes employee selection in that it shows that it is possible to control for moral hypocrisy. By making self awareness salient participants’ behaviour was aligned with their moral standards thereby eliminating moral hypocrisy.

The coin toss condition in the Batson et al. (1999) series of experiments allowed for ambiguity in that participants were able to pretend that the task assignment decision depended on the outcome of the coin toss. This ambiguity allowed participants to appear to behave in a moral manner while not being prepared to pay the cost of so doing, even when the issue of morality was made salient in the experiment of Batson et al. (1999). To investigate whether participants alter their behaviour in line with their standards (moral hypocrisy) or, alternatively, alter their standards to align with their behaviour (moral integrity). Batson et al. (1999) conducted a further experiment to see if, as Duval and Lalwani (1999) found, moral standards are made salient prior to the opportunity to behave that behaviour will be aligned with these moral standards. They eliminated the coin toss option. In doing so they found that participants in the low-standard-salience/high-self-awareness condition responded very differently to those in the high-standard-salience/high-self-awareness condition. In the latter condition the majority of participants agreed that the most moral way to assign the tasks was to give the positive consequences task to the other participant, whereas in the low-standard-salience/high-self-awareness condition

only a small minority agreed with this. This finding also has great relevance to the methodology used in the field research for this thesis. It shows that it is not unreasonable to expect that in the context of high stakes personnel selection, faking good, a form of moral hypocrisy, can be greatly reduced by following a procedure that mimics the procedures that Batson et al. (1997) procedures followed in their experiments. Therefore the importance of the Batson et al. (1997) and Batson et al. (1999) experiments, with respect to the research undertaken for this thesis, lies primarily in the fact that their research findings arguably support the use of a formal warning about the consequences of faking good before completing the personality in order to make moral standards salient for the participants. Moral standards were made salient by Batson et al. (1997) by including in the information sheet provided to participants the statement that *“Most participants feel that giving both people an equal chance – by, for example, flipping a coin – is the fairest way to assign themselves and the other participant to the tasks”* (p. 528). The purpose of this statement is analogous to that of the formal verbal warning used in the field study. In addition, the Batson et al. (1999) experiments showed that a combination of making moral standards salient together with heightened self-awareness eliminated moral hypocrisy. These experiments shed light on the on the psychological processes that inform theory building with respect to understanding the manifest faking good behaviour.

It is also arguable that the process of completing a self-report omnibus personality measure would lead to heightened self awareness in participants just as the simple act of seeing one’s reflection in a mirror did in the Batson et al. (1999) experiments. Duval and Wicklund’s theory of self-awareness (Duval & Lalwani, 1999) posits that focussing one’s attention on the self induces a state of objective self

awareness, and this leads to an awareness of the discrepancies between the ideal and actual self (Higgins, 1987). For objective self awareness to manifest itself the individual engages in introspection and self-evaluation while the individual, at same time, ignores endogenous environmental factors (Silvia & Duval, 2001). Self-awareness can be experimentally induced by exposing participants to self-focusing stimuli (Morin, 2011). The important aspect of 'objective self awareness' theory in understanding the phenomenon of faking good is the degree that a person's attention is focused upon a salient within-self discrepancy (e.g. perceived self evaluated discrepancy between actual and ideal, or ought, self in high stakes personality assessment) and provided that attention cannot be directed elsewhere (such as on moral values or standards because of low saliency), there will be efforts to reduce that discrepancy by faking good (Pryor, Gibbons, Wicklund, Fazio, & Hood, 1977). An individual completing a self-report personality measure is focussed on the self, by definition. This focus on, and real time awareness of, the discrepancies between the actual and ideal self in a setting such as that of a high stakes selection situation are arguably conducive to a state of moral hypocrisy among participants in the field study.

The low-standard-salience/high-self-awareness condition is commonplace in life, according to Batson et al. (1999), a circumstance they described as 'frightening'. Their concern arises from the fact that people are frequently asked to make moral decisions in circumstances in which the relevant moral standards are not stated in advance, or when others are watching, or when their actions are challenged, or when they do actually feel accountable for their actions, and so on. Therefore, according to Batson et al. (1999), many everyday moral decisions occur in low-standard-salience/high-self-awareness situations. High stakes employee selection contexts

would be a good example of such moral decision making situations. Therein lies the relevance and importance of the Batson et al. (1997, 1999) experimental findings to this field research and thesis. Individuals who fake good, when completing self-report personality measures in high stakes selection situations, could and probably would feel that they acted morally if questioned after completing the questionnaire, even though they actually acted in a manner that served their self interest. Individuals who are candidates for jobs are acting from a self interest perspective. They are seeking gains such as better employment and career opportunities, increased income, or some other such long and short term economic benefit. In addition, they are in competition with others for the positions on offer.

The setting in which participants completed the personality measure in this field study, and the procedure followed, make the field study analogous to that of the participants in the Batson et al. (1997, 1999) experiments. From this perspective the field study is in many respects an applied test of the findings of the Batson et al. (1999) Study 3 experiment. In recent years a number of other researchers, including Mazar, Amir, and Ariely (2008) and Shu, Gino, & Bazerman, 2011, have expanded on the work of Batson and colleagues.

4.2.2 Mazar's Dishonesty Research Paradigm

Mazar, Amir, and Ariely (2008) investigated the extent to which people who think highly of themselves in terms of honesty make use of various mechanisms that allow them to engage in a limited amount of dishonesty while retaining positive views of themselves. This is similar to Batson's moral hypocrisy concept. They used a novel experimental paradigm in their investigations. Specifically, the task that participants

had to complete consisted of two sheets of paper - a test sheet and an answer sheet. The test sheet consisted of 20 matrices, each based on a set of 12 three-digit numbers. Participants had four minutes to find two numbers per matrix that added up to 10. This is a straightforward search task, and though it can take some time to find the right answer, when it is found, the respondents could unambiguously evaluate whether they had solved the question correctly (assuming that they could add two numbers to 10 without error), without the need for a solution sheet. The answer sheet was used to report the total number of correctly solved matrices. At the end of the experimental session, two randomly selected participants would earn \$10 for each correctly solved matrix. In the control condition, at the end of four minutes, participants handed both the test and the answer sheets to the experimenter, who verified their answers and wrote down the number of correctly solved matrices on the answer sheet. In the experimental conditions participants indicated the total number of correctly solved matrices on the answer sheet, folded the original test sheet, and placed it in their belongings, thus providing them an opportunity to cheat. In a variation on the basic experiment participants were paid for each matrix solved regardless of whether the number correct was scored by the experimenter or reported to the experimenter by the participant. In one of the experiments the attention of participants was drawn to moral standards by giving the participants two minutes to write down as many of the ten commandments as they could remember. The same was also true when participants were reminded of the university's honour code and asked to sign a statement which read, "I understand that this short survey falls under [name of the university that each participant was attending] honour system." Participants were required to print and sign their names below the statement, before

they started answering the matrix problem questions, which was placed at the top of the answer sheet.

Mazar et al. (2008) found that when people had the ability to cheat, they cheated, but the magnitude of dishonesty per person was relatively low relative to the possible maximum amount, and that the level of dishonesty dropped when participants paid attention to honesty standards. Making moral codes salient reduced cheating completely. Another interesting finding was that even though participants knew that they were over-claiming their actions it did not affect their self-concept in terms of honesty. There have been a number of studies by other researchers into the question of moral hypocrisy which used the research paradigm of Mazar et al. (Barkan, Ayala, Gino, & Ariely, 2012; Ruedy, Moore, Gino, & Schweitzer, 2013; Shu, Gino, & Bazerman, 2011; Shu, Mazar, Gino, Ariely, & Bazerman, 2012). There was a slight difference in the research paradigm used to that of Mazar et al. (2008) in that in the cheating condition participants put their completed answer sheet into a shredder, and wrote down the number of matrix problems solved on a separate sheet which they handed to the experiment administrator. The shredder only made the sounds of shredding without actually shredding the answer sheet which had a unique participant identifier on it.

In further studies using the research paradigm of Mazar et al. research carried out by Shu et al. (2011) found that having participants read an honour code reduced cheating by half in the Mazar et al. (2008) experiment, by having participants read and sign an honour code almost completely eliminated cheating. In the no honour code condition 57% of participants cheated, in the read honour code condition 32% of participants cheated, and in the read and sign honour code condition 5% of participants cheated. Similarly, Chambers, Epley, Savitsky, and Windschitl (2008)

found that a slightly dim room increased cheating above and beyond the effect of guaranteed anonymity. Consistent with these findings, Shu et al. (2012) showed that signing an honour code statement before completing the matrix solving problems eliminated cheating whereas signing after completion of the task had little or no effect on cheating. The experiment included a simulated tax form completion exercise with an opportunity to cheat such that participants could cheat on the tax return form and get away with it by overstating their 'income' from the problem-solving task, and by inflating the travel expenses they incurred to participate in the experiment. The number of cheaters was lowest in the signature-at-the-top condition (37%), higher in the signature-at-the-bottom condition (79%), and somewhat in between those two but closer to the latter for the no-signature condition (64%).

The research of Shu et al. (2012) also included a field study based on the research paradigm used in the laboratory experiments which looked at the effect of the signature location on a self-report insurance policy review form in a naturalistic setting. Customers of an American insurance company are required to self-report the milometer reading of their car. The lower the miles driven the lower the accident risk, and therefore the lower the insurance premium. Customers asked to sign an 'honesty' statement at the top of the form had a reported mileage greater, on average, than those who were required to sign at the end of the form.

In sum, these experiments based on the Mazar et al. (2008) experimental paradigm, together with the field study, show that moral disengagement and moral hypocrisy occur under circumstances in which the environment is permissive i.e. no checks on the occurrence or the extent of cheating, and the absence of moral saliency. The research shows that once people begin to behave dishonestly by cheating, of which one form is lying, they disengage morally and will continue to cheat and lie in

permissive situations where there is an opportunity for self interest gain. On the other hand, it is relatively easy to prevent this moral disengagement and moral hypocrisy by simple environmental nudges. Warnings before job applicants complete self-report personality measures fall into a similar category in that the warnings are a form of simple environmental nudge towards morally appropriate behaviour. Without warnings about the consequences of faking good the environment in which the self-report personality measures are completed becomes permissive, particularly in the case of high stakes employee selection situations. The figures for the percentage of participants who cheat in the Mazar et al. (2008) experimental paradigm support the findings from Batson's research, and provide indirect evidence that the figure of 30% with a margin of error of + or – 10% for the incidence of faking good in job applicants selection situations found by Griffith and Converse (2012) is reasonable. More recently Lönnqvist, Irlenbusch, and Walkowitz, (2014) used the Mazar et al. (1995) experimental research paradigm to examine the question of whether the moral hypocrisy was a form of impression management directed towards impressing others rather than an intrapsychic phenomenon involving self deception. In the experiment participants could either directly choose a distributively fair (50/50) or selfish (80/20) allocation of money. Lönnqvist et al. (2014) found that the moral hypocrisy that occurred was a conscious attempt to impress the anonymous other participant or an unknown experimenter, and not by a primarily self-deceptive process aimed at sustaining one's self image as a moral person. They state that "we showed that the impression management of moral hypocrites may generally not be accompanied by self-deception" (p. 60). Research from the field of behavioural economics the research finding on moral hypocrisy in the previous two subsections, and is next reviewed.

4.2.3 Behavioural Economics and Moral Hypocrisy

Behavioural economists have also investigated lying and honesty in a range of experimental settings, using games in which players can announce future moves or can reveal (not verifiable) private information. This body of research has also been neglected by those I/O psychologists who have studied faking good in employee selection situations. Yet, just as faking good is an exemplar of moral disengagement in a particular context, experiments that study human behaviour in economic settings are also informative with respect to moral disengagement and hypocrisy, and also shed light on the question as to whether faking good in high stakes employee selection situations raises questions about generalisations with respect to the construct validity of personality measures. Croson and Sundali (2005) collected experimental evidence indicating that people depart from randomness in situations, and that lying behaviour that can occur is not in accordance with the prediction of the standard economic models. They found that people lied when this increased their profit. Fischbacher and Föllmi-Heusi (2013) developed a new and simple experimental design that makes it possible to detect lies when participants face no threat of being caught individually. This is arguably analogous to the situation of an individual completing a personality measure with no procedural, and/or statistical controls, for detecting or preventing impression management.

The experiment is a one shot single decision making situation, and the following description is taken directly from the Fischbacher and Heusi (2008) paper in which the procedure was first published. It took less than ten minutes to conduct the experiment. Participants were each provided with a six-sided dice. They were

informed not to touch the dice until requested to do so. The experimenter then told the participants that instructions would be given on the screen. Participants then read these instructions and were informed that they were going to receive a payoff for filling in a questionnaire, and that this payoff would be different for each participant. To determine their individual payoff, the participants were requested to roll a dice. The payoff would equal 1, 2, 3, 4, and 5 Swiss Francs (CHF) if the dice came up with the corresponding number and zero CHF if the dice came up with a 6. Participants were explicitly called to roll the dice more than once in order to check whether the dice was fair. It was highlighted on every screen that only the first throw was relevant for the payoff and therefore should be kept in mind. Next, participants were requested to roll the dice and to memorise the figure rolled. On the last instructing screen, participants reported the number rolled together with the resulting payoff. In this experiment, lying means reporting a different number than the one actually rolled in the first throw. It was impossible to detect lying on the individual level.

The findings of studies using this research paradigm showed that 20% of participants lie to the fullest extent possible while 39% of subjects are fully honest. In addition, the remaining participants consists of partial liars in that these participants also lie, but do not report the payoff-maximizing roll of the dice (Fischbacher & Heusi, 2013). Pruckner and Sausgruber (2013) subsequently published the results of an interesting Austrian field study which are consistent with the findings of Fischbacher and Heusi. The results of the field study experiment examined the role of honesty norms among customers in a real market for newspapers where payments are not monitored. Austrian print production companies commonly sell tabloids on the streets, via an 'honour system' that involves a booth containing a bag filled with newspapers and a padlocked cashbox. Customers are supposed to deposit payment

into the cashbox, but this payment method does allow them to underpay or simply take the paper without paying, if they so choose. The booth contained a message – either ‘stealing a paper is illegal’ or ‘thank you for being honest’ - for customers that reminds them of the moral implications of their action with respect to paying for the newspaper. Overall the study showed that 39% of customers paid nothing, 42% made a payment that was below the price of the paper, and 19% paid the full price.

These results mirror the findings of Zickar, Gibby, & Robie (2004) that used the mixed method item response theory technique for their data analysis of responses by job applicants and incumbents to items in a short form personality measure accompanied by a written warning which stated that distorted self-descriptions would invalidate the respondents’ test results. They found that there were three classes of applicants needed to model all responses patterns – regular responders who didn’t fake, slight fakers, and extreme fakers.

The extant research reviewed in Section 4.2 shows that there is a consensus that faking good can, and does, occur in occupational settings. The review also shows that there is a considerable body of research which shows that moral hypocrisy is a ubiquitous phenomenon, and that moral hypocrisy in the form of faking good is a real issue in ‘high stakes’ employee selection situations (Griffith & Converse, 2012; McFarland, 2003). This factor must be included in the nomological net (Cronbach & Meehl, 1955) underlying personality assessment in general. In particular, from the perspective of this research programme, in high stakes employee selection situations remedies for dealing with its possible occurrence must be considered and evaluated.

4.3 Remedies for Dealing with Faking Good

A number of possible remedies have been suggested for dealing with the problem of faking good (Salgado, 2016). These include the use of Social Desirability or Lie Scales either administered separately or by means of a validity scale that is included in the personality inventory. Examples of the former include the Marlow-Crowne measure (Crowne & Marlowe, 1960) and the Balanced Inventory of Desirable Responding (Paulhus, 1984). The Minnesota Multiphasic Inventory (MMPI) which is used mainly in clinical settings includes a 'Lie' scale as does Eysenck's EPQ measure (Paulhus, 2002). The California Psychology Inventory (CPI) has a 'Good Impression' scale. Of eleven popular commercially available personality inventories Goffin and Christiansen (2003) surveyed eight of them included some form of social desirability scale. Even though the use of such scales has been frequently questioned (Dilchert & Ones, 2012). Goffin and Christiansen (2003) found that 56% of practitioners in applied psychology who they surveyed used social desirability scales of one form or another to correct scores on personality measures.

Other methods that are used for dealing with faking include Response Inconsistency Scales. These scales are purported to work as a method for detecting faking by means of the principle that certain response patterns to particular item pairs in questionnaires consisting of different, yet similar, items are inconsistent if the item pairs are carefully matched (Dilchert & Ones, 2012). Response latency to items in personality measures is another method used. The principle underlying this approach is that individuals are more likely to give a socially desirable answer if they had plenty of time to respond to items in a personality inventory (Robie, Komar, & Brown, 2010). However, Robie, Curtin, Foster, Phillips, Zbylut, and Tetrick (2000)

found that individuals who had been coached on how to beat response latencies could, in fact, ‘beat’ the response latencies, but could not then elevate their personality scores in comparison to a group of participants instructed to answer honestly.

Vasilopoulos, Reilly and Leaman (2000) found in laboratory studies with student participants that faster responses were linked to impression management for those jobs that the participants were familiar with. However, this did not hold for unfamiliar jobs where impression management was indicated by slower response times. Dilchert and Ones (2012) maintain that, with respect to response latency measures, “there is at present no compelling data supporting their widespread use in noncognitive assessments” (p. 187).

At the item level a number of strategies are used in order to minimise impression management or faking. Item placement, reverse scoring, the use of subtle versus obvious items, and forced choice formats have all been looked at. All of these item level strategies are designed to obstruct the test taker in identifying the constructs being measured. The empirical research on these topics is sparse (Dilchert & Ones, 2012). However, there is some evidence that test takers do look for patterns among the items (Weijters, Geuens, & Schillewaert, 2009).

The use of forced choice questions has been long debated. However, there is a major technical problem with this item response format compared with items that are scored using a Likert scale. This is a problem that is inherent to ‘ipsative’ scoring. The term ipsative is used roughly as a synonym for ‘interdependent’, and refers to some type of dependency among the variables measured on a given survey (Meade, 2004). In a typical ipsative personality test used for employee selection items are paired or grouped together in an item set. According to Meade (2004), “ipsative measures will be extremely inefficient for use in employee selection” (p. 548). Adair

(2014) in two meta-analyses found that warnings are generally more effective than forced-choice or item transparency interventions at reducing faking behaviour, and that randomising the order of items does little to influence faking. He also showed that interventions, such as warnings, to deal with faking are generally effective at reducing faking behaviour, as evidenced by smaller sample-weighted mean effect sizes for studies with a faking intervention compared to those without any intervention.

Rothstein and Goffin (2006) maintain that the most effective way to limit the effects of faking is to employ a faking warning. McFarland (2003) showed that warnings in the form of statements that caution test takers against deliberate response distortion are a good strategy to use in dealing with faking good when administering personality measures in selection situations because such warnings reduce multicollinearity between the personality dimensions. In an earlier study Kluger and Colella (1993) also found that warning against faking does reduce faking behaviour. Their findings show that warnings reduced the extremeness of the item means and increased item variability for scales composed of mostly obvious or transparent items in regard to job desirability, which was defined by the authors as presenting oneself as possessing qualities that are perceived to be important for the particular job. Dwight and Donovan (2003) pointed out that “Given that applicant faking appears to be a valid concern, it follows that such faking needs to be combated in some manner” (p. 2). They found that those who received a warning scored lower on the personality scales than those who did not receive a warning. Dilchert and Ones (2012) point out that one advantage of warnings is that they can be easily implemented when personality tests are being administered at little or no cost. They also state that “it would certainly be encouraging if a simple statement that faking could be detected

(and would have negative consequences, whether true or not) would result in a reduction of its prevalence” (p. 193). Salgado (2005) makes the point with respect to the use of a formal warning that “this strategy is both valid (it reduces distortion) and economical” (p. 123).

In reviewing the different types of warnings from Pace and Borman’s (2006) taxonomy Dilchert and Ones (2012) strongly advise that the common pitfalls of over reliance on laboratory studies and unrealistic experimental conditions that fail to stimulate real-world incentives to distort be avoided. The Pace and Borman taxonomy provides a useful review of the different kinds of warning that can be used in the administration of personality measures in employee selection situations. These are warnings about detection based on simply claiming that detection methods are included in the assessment process, warnings about the consequences of faking which can range from mild (retesting) to severe (elimination from the applicant pool), appeals to reason which point out the importance for the applicant of obtaining an accurate profile, appeals to moral principals pointing out that faking is an example of immoral test behaviour, and warnings of an educational nature about why truthful responding is important in order to obtain accurate results.

According to Gilovich, Savitsky and Medvec (1998), when people are aware that their lies can be detected the phenomenon of the ‘illusion of transparency’ occurs. Liars feel as if their feelings of nervousness about lying can leak out, or that others could "see right through them" (p. 335). In the studies carried out by Gilovich et al. (1998) participants who were induced to lie overestimated the detectability of their lies. This research show that when there is a possibility of lie detection occurring the illusion of transparency can happen, and is robust across a variety of procedural changes in the domain of lie detection. People have more information about

themselves than others have. There is a fundamental asymmetry in what people know about themselves compared to others. This helps to explain why individuals have difficulty accurately intuiting how they appear to other people (Chambers, Epley, Savitsky, & Windschitl, 2008).

From a construct validity perspective, this section combined with Sections 4.1 and 4.2 clearly show that socially desirable responding in the form of faking good is an issue when it comes to accurately assessing the Big Five in high stakes employee selection situations. This must be taken into account when making inferences particularly about the rank order of candidates in high stakes selection situations. The remedy used in this research programme was the formal warning procedure. The next section draws together the research reviewed in the previous sections of this chapter.

4.4 Conclusions from the Research Reviewed

The review of research from behavioural economics, as well as that of Mazar et al. (2008), Shu et al. (2011), and others dealing with moral hypocrisy, lends very strong support to those I/O psychologists who argue that there is robust evidence that supports the construct invalidating role of faking good or impression management, when it comes to inference drawing regarding the suitability of some job candidates. Consequently, there is a need for both procedural and statistical controls (Podsakoff et al., 2003) to minimise the incidence of false positives in the selection and recruitment procedures of organisations arising from faking good. This factor forms part of the nomological net with respect to establishing the construct validity of personality

assessment that Cronbach and Meehl (1955) emphasised in their seminal paper on the topic.

In contexts which are lacking in clear guidelines or norms of behaviour, and in which self-interest is pitted against being honest, ambiguity can serve as a justification to do wrong by faking good for instance, but the person still feels that he or she is a moral individual (Pittarello, Leib, Gordon-Hecker, & Shalvi, 2015). That is, people's attention is more easily shifted toward tempting information in ambiguous settings than in unambiguous settings, and this tempting information then shapes their self-serving lies. Pittarello et al. (2015) recommend that by crafting environments in which ambiguity is low and transparency high the temptation to engage in moral disengagement will tend to push individuals toward less moral hypocrisy, and cause them to stick more to the ethical standards they would purport to hold and abide by, if questioned.

Self-serving justifications determine the extent to which people stretch the truth. For example, it may well be that the availability of coaching in how to beat psychometric tests leads job applications to the self-justification belief that it is morally acceptable to engage in faking good. There are companies that specialise in training job candidates in interviewing skills and in CV preparation (Sliter & Christiansen, 2012) such as, for example, "Make a smart investment in your career – get a professional psychologist to coach you to ace the psychometric test" (Institute of Psychometric Coaching, 2017). As Sliter and Christiansen (2012) point out all of these services engender social norms of how to present in high stake job selection situations in a manner that is designed to win the position on offer for the applicant rather than ensuring a job/application fit that is good for both the potential employer and the applicant. Potential employers have long used techniques such as reference

checking and work samples to try to get an accurate picture of candidates arguably because this approach can help to counter moral hypocrisy on the part of job applicants.

As already described in Section 4.1, Hogan, Barrett, and Hogan (2007) are to the fore in arguing the case against faking affecting the construct validity of personality measures in the assessment of job applicants in high stakes selection situations when they state that “Results suggest that faking on personality measures is not a significant problem in real-world selection settings” (p. 1270). They do, however, acknowledge that it is a problem while maintaining that it is not a significant one. Yet all of the research evidence from the nomological net (Cronbach & Meehl, 1955) that surrounds the construct of moral hypocrisy suggests that faking good or impression management is, indeed, a significant problem. The evidential opposite case put forward by Griffith and Converse (2012) that roughly 30% of applicants engage in faking behaviour is much more robust than the evidence of Hogan et al. (2007). Griffith and Converse’s (2012) evidence is strongly supported by the research presented from evidence on the topic of moral hypocrisy and behavioural economics detailed earlier.

When the validity of a personality measure is being discussed it should be borne in mind the important difference in work and organisational contexts between aggregated anonymous personality assessments, that are used for the purposes of establishing the concurrent and/or predictive criterion related validity of the Big Five dimensions of personality, and the construct validity of the same Big Five measures when used for the purposes of selecting the ‘winner’ in high stake job selection situations (Embretson, 1983; Messick, 1995; Sackett, 2012). There is an economic payoff to be gained in the latter situation – either a promotion with higher status and

remuneration, a change of employer to a more putatively desirable one, entry into the workforce, etc. All of these desired outcomes, for the applicant, have a potential payoff attached that is frequently monetary. For this reason alone the research on moral hypocrisy (Batson et al., 1997; Batson et al., 1999; Mazar et al., 2008; Shu et al., 2011) is of great relevance, and very informative, with respect of the extent to which lying, euphemistically labelled as impression management or faking good by I/O psychologists, does occur in high stakes employee selection situations.

As mentioned earlier, Griffith and Converse (2012) put the figure for the extent to which such lying occurs at 30% with a ‘confidence interval’ of + or – 10%. Indeed, it is arguable that an upper limit of 40% or even higher is supported by the research on moral hypocrisy. The research reviewed in this section has shown that lying, cheating, or moral disengagement is a heterogeneous phenomenon that is situationally dependent. In any given situation where ambiguity can potentially lead to the occurrence of moral hypocrisy even though some don’t lie or cheat at all, while some lie or cheat to the maximum extent possible, and the remainder lie or cheat to an extent. Pruckner and Sausgruber (2013) in their newspaper purchase field study found that only 19% of participants paid the full price. Fischbacher and Heusi (2013) found that only 39% of participants in their study were fully honest. In the no honour code condition of the Shu et al. (2011) studies 57% of participants cheated. In the Batson et al. (1997) study of those who choose to make the decision based on the toss of the coin 85% to 90% assigned themselves the interesting and rewarding task, even though an analysis based on chance says that the figure should have been around 50% for those who tossed the coin.

In light of research findings from both the study of faking behaviour in personality assessments by I/O psychologists, and the research from other fields of

research described earlier, it is difficult to argue for a nomological net which endorses the position that faking good does not occur and/or is not an issue of construct validity concern, as Hogan et al. (2007) did, when job applicants in high stakes selection situations complete personality assessments using self-report measures. The methodological and applied consequences of this rejection of the Hogan et al. (2007) position is that a) research into the underlying hierarchical structure of the Big Five can be contaminated by method variance and this should always be investigated and catered for, and b) the ranking of applicants will be biased in the absence of procedural controls to minimise, if not eliminate, this bias. As a result of these two factors the objective of using high performance work practices will not be achieved in those organisations using personality assessment as part of their selection and recruitment processes and procedures.

Job applicants' pre-testing justifications for faking good have been shown to be sensitive to interventions that eliminate ambiguity (Shu et al., 2011). This is the approach taken in the field research used in the research for this thesis. One such intervention is, as discussed in Section 4.3, through the use of a pre-test formal warning which makes clear and clarifies in unambiguous terms the ethical conduct expected of candidates. This makes morality salient as Batson et al. (1999) and Shu et al. (2012) showed, because procedural interventions such as drawing attention to an honour code, that increase the salience of a specific ethical code, have been shown to be effective in preventing the occurrence of moral disengagement and moral hypocrisy as described in detail earlier. Ethical salience intensifies the threat to the self, and decreases the power of justifications for the moral disengagement and unethical behaviour. This is, arguably, the role of warnings to candidates prior to completing a personality measure.

This review has shown the issue of faking good on personality measures in employee selection context, as well as that of moral hypocrisy, has been extensively looked at in research using both laboratory studies and field studies (Batson, 2008; Griffith, & Peterson, 2011; Rothstein and Goffin, 2006, Viswesvaran & Ones, 1999). It can be concluded from the review that, from a construct validity perspective as set out in Chapter 2, questions remain to be answered as to the construct validity of personality assessments in high stake employee selection situations. In order to establish a test as a construct valid measure of a latent construct the underlying nomological network and theory must be sufficiently established, so that others can accept or reject it (Cronbach & Meehl, 1955; Loevinger, 1957; Embretson, 1983; Messick, 1995). Without the procedural precautions of a formal warning to eliminate or minimise faking, and the use of a construct valid impression management measure, it is arguable that the inferences made about individuals assessed using personality measures in high stakes selection are not construct valid. An additional problem with the research is that there is no agreement on how to measure faking good (Burns & Christiansen, 2011; Griffith & Peterson, 2008; Marcus, 2009; Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). These two construct validation issues are what the research programme of this thesis had to address in the field research.

In summary, faking good does occur in selection situation and the extent of to which it does is alarming from the perspective of construct validity. The issue of construct validity, and the rank order of candidates when it comes to unfairness and bias in high stakes employee selection situations, is an important one. Failure to take this into account is inconsistent with what Messick (1995) and others (Embretson, 2007; Kane, 2001; Smith, 2005; Strauss & Smith, 2009) recommend for the establishment of the scientific standing of the construct validity of a measure. The

review also shows why Campbell and Fiske's (1959) MTMM approach to separating trait variance from method variance can shed light on the hierarchical structure of the Big Five. Job applicants can, and do, fake good. The issue of fairness in testing, the possibility of selecting executives that will derail at some time in the future due to lacking the necessary competencies (Hogan & Hogan, 2001), and/or the dangers of sub clinical narcissism and psychopathy, the extent to which fakers are likely to be in subsets of finalists, and/or selected in executive selection are important. These issues all need to be examined as part of the process of establishing the construct validity of the NEO PI-R in high stakes employee selection situations.

One research method by which the issue of the measurement of faking good on personality trait scores has been addressed in research is by examining the effect size of the difference in participants' scores on the personality dimensions. Job applicants' and job incumbents' scores are compared regardless of whether the research is conducted on a between participant or within participant basis (Dilchert & Ones, 2012). This approach is not a feasible one in the applied setting of job candidate selection contexts, particular when it comes to selecting a single candidate from a shortlist. The most widely used alternative method has been to use a stand alone social desirability scale such as the IM scale of the BIDR of Paulhus (1998) - which is the approach taken in this field study – or by means of a bespoke socially desirable scale embedded in the omnibus personality inventory (Connelly & Chang, 2013; Hopwood & Donnellan, 2010). This is an important topic and will now be examined in some detail in the next chapter.

Chapter 5

Accounting for Impression Management

The previous chapter reviewed the background research on the effect of socially desirable responding on the measurement of the Big Five using self-report measures. This chapter contains a review, in Section 5.1, of the current status of the measurement of socially desirable responding using what are broadly referred to as ‘lie scale’ measures. These measures are used to putatively detect faking good (or faking bad in clinical contexts) in self-report measures, as well as the investigation of aspects of the internal factor structure of such lie scales in general which has been a source of debate with respect to the measurement of socially desirable responding (Paulhus, 1984). Secondly, in Section 5.2, the research evidence that bias and unfairness due to faking good can occur in employee selection decisions is examined. This is a necessary requirement because of possible consequential rank order effects in high stakes selection contexts when faking good which occurs is not accounted for. This issue is of critical importance to the objectives of this research programme. The approach of Messick (1995) to construct validity was broadly followed. It is important to keep in mind when reading this chapter that, because most faking good measures in use are also self-report measures, some of the contents of the previous chapter are also relevant to the topic of this chapter.

A valid observation (Kunchel, Borneman, & Kiger, 2012) concerning all self-report measures is that “One person’s deceptive response looks the same as another’s honest self-report” (p. 104). This presents a problem when trying to devise a measure

to detect impression management. It has been shown in Chapter 4 that there is a broad consensus that common method variance (CMV) due to faking good is a major concern when it comes to personality assessments in high stakes employee selection situations. Faking good and impression management are forms of socially desirable responding which are not easy to detect. There is also the question of whether impression management is a conscious or subconscious phenomenon (Paulhus, 1984) to be considered. If, in spite of taking procedural precautions such as using a formal warning, the procedure followed in this research programme did not take fully account of the occurrence of faking good among job applicants (if not eliminated by the use of a formal warning) then the criteria for establishing strong construct validity will not be met (Messick, 1995; Embretson, 2007). Some of the inferences made, based on an individual candidate's scores on the measure of the Big Five used, may still not be valid in spite of the formal warning.

Impression management in the form faking good must be fully accounted for, if at all possible. A number of approaches were described in Section 4.3 of the previous chapter for dealing with the problem. The method of accounting for socially desirable responding adopted in this research programme is examined in some detail in this chapter. Given the use of the NEO PI-R personality measure and its Likert scaling approach, the question of using a forced choice items approach (Drasgow, Stark, Chernyshenko, Nye, Hulin, & White, 2012) did not arise.

5.1 Lie and Related Scales

It is clear from the literature that faking good and impression management have been a cause for concern for decades (Paulhus & Reid, 1991; Paulhus, & Vazire,

2007; Uziel, 2010) when dealing with self-report measures of latent psychological constructs. Consequently, in order to try to deal with the issue of faking a number of measures that purport to detect and measure such socially desirable responding, which are themselves self-report measures and therefore also questionable from a construct validity perspective, have been used over the years.

The construct validity of these measures has been hotly debated for many years (Block, 2010; Burns & Christiansen, 2011; Costa & McCrae, 1997; Dilchert & Ones, 2012; MacCann, Ziegler, & Roberts, 2012; Paulhus, 1984). In fact, the use of lie scales of any form to detect faking in self-report measures has been questioned by some researchers who go so far as to recommend against using lie scales as a method for detecting faking good (MacCann et al., 2012). So the issue of the construct validity of measures of socially desirable responding as well as the validity of the inferences of the personality measure used is a very important issue in evaluating the research methodology used in this thesis. This is because the method that was used to detect faking good in this field study relied on a bespoke version of one particular measure – Paulhus’s Balanced Inventory of Desirable Responding Impression Management measure (Paulhus, 1984) - from the range of lie scale measures that are in use.

There are two categories of such measures in use – unidimensional lie scales which assume that there is a single latent factor underlying the lie scale, and that of Paulhus (1984) which posits that there are two latent factors.

5.1.1 Unidimensional Lie Scales

In 1930, Hartshorne and May developed a lie scale to detect and help deal with socially desirable responding. High scores on the lie scale were assumed to be indicative of a dishonest character (Paulhus, 2002). Later on, in clinical settings, the widely used omnibus Minnesota Multiphasic Inventory (MMPI) included an embedded socially desirable responding scale, the MMPI Lie Scale, designed to identify individuals deliberately dissembling their clinical symptoms (Hathaway & McKinler, 1989). Later on the Eysenck Personality Inventory (EPQ) came into use and it, too, contained a Lie Scale (Eysenck, 1968). Additionally, two widely used stand-alone socially desirable responding detection scales are also widely used (Paulhus, 2002), namely, the Marlowe Crowne SD Scale (MCSD) and the Edwards SD Scale (ESD). These latter two measures contained items claiming improbable virtues and denying common human frailties. High scores were accumulated on these socially desirable responding measures from self-descriptions that were not just positive, but improbably positive. Typical items include the following - “I always try to practice what I preach” from the MCSD, and “No one cares much what happens to you” from the ESD (Shaver, Brennan, Robinson, Shaver, & Wrightsman, 1991). A number of the omnibus personality measures in use today such as for example Eysenck’s EPQ and the 16PF have the lie scale embedded in the personality measure (Ellingson, Smith, & Sackett, 2001). These embedded measures have been the most widely employed applied technique to deal with applicant faking (Barrick & Mount, 1996; Holden, 2007; Hough, 1998; Hough, Ones, & Viswesvaran, 1998; Kurtz, Tarquini, & Iobst, 2008; White, Young, Hunter, & Rumsey, 2008), particular in applied settings. It should be noted at this point that the NEO PI-R, which was used

in the field study for the research programme of this thesis, does not have an embedded lie scale.

As mentioned above some of these lie scales have been used extensively in research on faking good and are also widely used in applied settings (Holden & Book, 2012). For example, a study using four different lie scales, some embedded, by Ellingson, Smith, and Sackett (2001) claimed to show that the factor structure of personality was invariant when comparing the two groups of participants that were investigated in the research – those who scored high on the impression management measures and those who scored low on the same measures in the four samples included in the study. However, according to Ellingson et al. (2001), construct validity was negatively affected due to inflated scores by participants who scored high on the lie scales.

In research and applied settings elevated scores on the lie scale used are taken to mean that the respondent was dishonest in the assessment (Hough, 1998). Scores on these measures are then used by some researchers and practitioners to supposedly partial out the variance in personality responses associated with faking in an attempt to obtain more accurate estimates of the criterion validity of personality tests (Smith & Ellingson, 2002). However, this approach has little empirical support as a valid technique for eliminating common method variance (CMV) due to faking (Dilchert & Ones, 2012; MacCann, Dilchert, & Roberts, 2012). Moreover, the unidimensional nature of lie scales was questioned by a number of researchers as far back as the 1960's, which led to Paulhus examining the factor structure of the lie scales in use (Paulhus, 2002).

5.1.2 Paulhus's Socially Desirable Responding measure

The factor analysis studies of Paulhus in the 1980's of the various socially desirable responding measures that were in use at that time found that there were two, rather than one, socially desirable responding factors which he referred to as self-deception enhancement and impression management, respectively (Paulhus, 1998). Self deceptive enhancement refers to an unconscious positive bias in item responses. It might occur when individuals complete the self-report measures with the aim of protecting positive self-esteem. In contrast, impression management refers to the conscious dissimulation of item responses with the aim of making a favourable impression on others (Paulhus, 2002). Sackett's (2012) multiple component analysis of systematic variance in responding to items in a self-report measure is consistent with this latent two factor structure of socially desirable responding of Paulhus (1984). Sackett (2012) differentiated between what he termed 'erroneous self-perception' and 'situationally specific intention' (p. 331) distortion. The former is an automatic response mode, whereas the latter is a controlled response mode. In the automatic response mode the test taker has a tendency to automatically respond to items in terms of her or his best self.

As a result of his research Paulhus (1984, 1998) developed a forty item measure – the Balanced Inventory of Desirable Responding (BIDR) - to measure the two factors of socially desirable responding. All of the forty items are affirmation statements, and there are equal numbers of attribution and denial items for each of the two 20-item sub scales measuring Self-Deceptive Enhancement (SDE) and Impression Management (IM). The BIDR is the most widely used stand alone socially desirable responding measure in both research and applied settings (Ellingson,

Heggestad, & Makarius, 2012). The IM scale of the BIDR has been widely used in the study of the effect of faking good on personality measures in occupational settings (Ellingson et al., 2012). This scale has also been used in research into the hierarchical structure of personality as a method for measuring impression management (DeYoung, Peterson, & Higgins, 2002).

Arguably Paulhus's (1984) distinction between the two dimensions of socially desirable responding can help to shed better light on the structural aspect of construct validity (Loevinger, 1957; Messick, 1995) of the Big Five, unlike other measures of socially desirable responding such as the unidimensional lie scales, which do not distinguish between these two dimensions, because of the volitional nature of IM compared with SDE (Lönqvist, Irlenbusch, & Walkowitz, 2014; Sackett, 2012). The findings of Lönqvist et al. (2014) support the contention that the BIDR-IM scale does assess the tendency to consciously, rather than subconsciously, give inflated self-descriptions to an audience. The research reviewed in Chapter 4 would also suggest that the BIDR-IM may indeed measure faking good because it is a form of intentional distortion similar to that observed in the research studies of Mazar, Amir, and Ariely (2008) and those of Shu, Gino, and Bazerman, (2011). A contrasting view is expressed by Uziel (2010) and will be explicated in the next section.

5.1.3 Construct Validity and the BIDR

The construct validity issue here is twofold. Are self-report measures of faking good truly construct valid? Secondly, even if they can be shown to be construct valid under what circumstances can faking good be accurately measured?

Recently, Uziel (2014) argued that the BIDR-IM measure was not associated with excessive self-enhancement and that those who score high on impression management, as measured by the BIDR-IM, are simply presenting a valid portrait of themselves. If Uziel (2014) were correct then the construct validity of the bespoke version of the BIDR-IM measure used in this research programme would be questionable. However, it is important to note that Uziel's (2014) evaluation of the BIDR was a correlational study which was not based on samples of participants who were dealing with moral dilemma type situations, such as that of the field studies of this research programme. This apart from anything else raises the question of the generalisability (Messick, 1995) of Uziel's (2014) findings, if they are correct, to high stakes employee selection situations. Nonetheless, the construct validity issues that Uziel's (2014) article raises have to be considered from the perspective of Cronbach and Meehl's (1955) nomological net and Embretson's (1983) nomothetical span. The nomological net refers to a system of interlocking statistical or deterministic laws, including some observables that are inter-related to some degree. The network may be incomplete because the chain of inferences underlying the construct is being developed and expanded upon. Embretson (2007) showed that construct validation is a dynamic process with various feedback loops. Therefore the nomological network underlying a latent can evolve as empirical evidence for and against the putative construct accumulates.

McFarland and Ryan (2000, p.818) proposed a model of the nomological network of faking that includes many of the factors that may interact to create variance in non-cognitive self report measures. Their model has many elements in common with the Theory of Planned Behavior (Ajzen, 2001). It meets Cronbach and Meehl's (1955) requirements for some form of evidence that there are a) observables

in the network with predicted relationships, and b) the relationships between the observables are ‘reasonably explicit’ (p. 300). A slightly modified version of this model is shown in Figure 6 to explicitly describe a putative nomological network for the Paulhus BIDR IM scale.

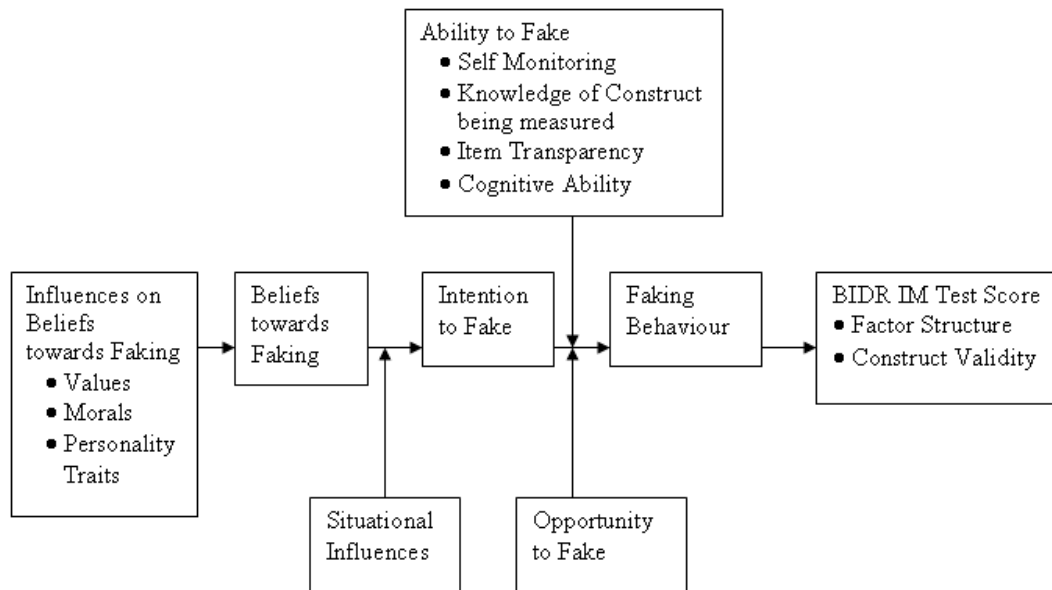


Figure 6 *Nomological Net for the Paulhus BIDR IM Scale*

Beliefs toward faking may influence faking behaviour. For example, some people may believe that faking is wrong, no matter what the circumstances are. If someone has a belief that faking is acceptable, then that individual is more likely to intend to distort his or her responses than is someone who believes faking is wrong (MacFarland & Ryan, 2000). The relationship is moderated by the extent to which individuals feel that they may gain a desired outcome such as getting a job by faking. The effect of beliefs toward faking on intention to fake is moderated by situational influences. The research programme of this thesis only examined three of the observables in the nomological net. Firstly, there was one issue arising from the network that was specifically catered for, namely, the issue of the Item Transparency

of the bespoke BIDR IM measure used in the research programme. Secondly, later on Table 20 in Chapter 8 shows that there was no correlation between the bespoke BIDR IM measure used and the two measures of cognitive ability included in the test battery that participants in the Managerial field study completed. Thirdly Chapter 6 contains a review of the background theory concerning the role of personality traits and situational influences in the empirical measurement of impression management.

There is an important issue concerning the use of the BIDR-IM measure in high stakes employee selection contexts in that the items in the IM scale are deliberately both overt and clear cut (Paulhus, 1998). This presents a possible construct validity problem in high stakes selection situations – presenting the twenty items as a single group to subjects in selection situations is a concern in this regard because of the overt and clear-cut nature of the items. Based on the research of McFarland, Ryan and Ellis (2002), into the effect of random item placement compared to item grouping in self-report questionnaires, it can be strongly argued that job applicants would be able to recognise the objective of the IM items when they are presented together as a single group of twenty items because of item transparency. The Batson et al. (1997, 1999) studies, which were reviewed in the previous chapter, show how easily moral hypocrisy can impact on an individual's willingness to pay the price of behaving in a morally honest manner. A recent study by Lonnqvist, Irlenbusch, and Walkowitz, (2014) is also important in helping to establish the construct validity of the BIDR-IM measure in that it showed that impression management as measured by the BIDR-IM scale can indeed be prone to the moral hypocrisy effect. Presenting the items in a self-report questionnaire as a single group can provide a strong environmental cue as to the objective of the measure, particularly in a high stakes selection situation. Arguably moral hypocrisy will likely lead to

faking good for some job candidates when the BIDR-IM items are all presented together.

Another construct validity concern about the BIDR-IM scale and self-report measures more generally including personality measures is the issue of pre-testing coaching. Sliter and Christiansen (2012) looked at the effect of self coaching on faking good on a personality measure and the BIDR-IM scale. Participants who read chapters from a commercially available book, 'ACE the Corporate Personality Test' (Hoffman, 2000), on how to 'beat' personality tests were more successful at distorting their responses than those participants who had not read the coaching book.

Individuals who self coached before completing a Big Five personality measure elevated their personality scores, primarily on the traits that had been targeted in the coaching. In addition, those who were also self-coached on avoiding lie detection scales scored significantly lower on the BIDR-IM scale while simultaneously increasing their personality scores. In another study, Robie, Komar, and Brown (2010) provided participants in their study with a coaching video. Participants' exposure to this video resulted in a similar level of score elevation to that found by Sliter and Christiansen (2012) on the dimensions of the Big Five compared to when the same individuals had no training and who were more likely to be responding honestly. However, unlike Sliter and Christiansen (2012) they found that the 15 minutes coaching video that they used had no effect on the BIDR-IM scores. The findings with respect to the effect of coaching on the BIDR-IM scores from these two studies are contradictory. This may be due to the different procedures used to provide participants with coaching – the Sliter and Christiansen (2012) study relied on self coaching, and the Robie et al. (2010) study used a time limited, more didactic method. So from a construct validity perspective, there are legitimate concerns about the

validity of the inferences that may be drawn from the use of the BIDR-IM scale as a means of detecting and measuring the incidence and extent to which faking good occurs. Just as is the case with personality measures insight into this issue can be gleaned from the research into moral hypocrisy, as mentioned earlier. Both the Mazar et al. (2008) studies and those of Shu et al. (2011) showed that by making ethical behaviour salient before an event that has scope for moral disengagement to occur, through knowing that some form of monitoring will occur, the incidence of moral hypocrisy is greatly reduced. This highlights the fact that the accurate measurement and detection of faking good can be context dependent. It follows that inferences questioning the construct validity of a measure such as the BIDR-IM scale may not always be generalisable across situations. Therefore, blanket rejection of the validity of all lie scale measures is not consistent with the unified modern understanding of construct validity which would include consideration of a number of aspects such as, for example, the placement of items in such measures and the specification of the testing conditions (Messick, 1995; Embretson, 2007).

A meta-analysis of studies that included the two BIDR factors by Li and Bagger (2006) showed that that scores on neither the SDE nor the IM measure had any 'spurious' effect on the criterion-related relationship between personality measures and performance. Furthermore, they did not function as performance predictors. The lack of an impression management effect, as measured by the BIDR-IM scale, on criterion-related validity was shown to be the same for all of the Big Five dimensions (Li & Bagger, 2006). These findings lend some support to the use of the BIDR-IM because they are consistent with the Ones and Viswesvaran (1998) finding of no criterion related validity effect of faking good, already referred to in Chapter 4.

All of the foregoing may explain why in their study of personality testing and re-testing for managing intentional distortion Ellingson, Heggstad, and Makarius (2012) and Fan, Gao, Carroll, Lopez, Tian, and Meng (2012) used the BIDR-IM to detect faking good in their research. This research programme builds on these earlier studies. Ellingson et al. (2012) defended the use of the BIDR-IM scale in their research (Ellingson et al., 2012) on the following basis that “We measured intentional distortion using a validity scale because the practice of retesting emerged as a technique for responding to validity scale scores. We were careful to choose a validity scale that is widely used and explicitly designed to measure score invalidity” (p. 1074). They used the short form NEO-FFM personality measure and re-tested all participants who scored above a BIDR-IM cut-off score. The effectiveness of retesting in Ellingson et al. (2012) study was very much a function of making valid inferences regarding the degree to which an individual is engaging in deliberate falsification of responses. Fan, Gao, Carroll, Lopez, Tian, and Meng (2012) used the BIDR-IM scale to detect faking good after participants received a formal warning. They found that job applicants do indeed engage in faking and that levels of faking were reduced, although not completely eliminated, after applicants whose scores on the IM measure suggested that they were faking were retested after receiving a targeted warning.

The final, and probably the most important, piece of evidence in support of the construct validity of the BIDR-IM scale comes from a recent paper of Connelly and Chang (2016). They carried out a meta-analytic confirmatory factor analysis MTMM study of personality traits, socially desirable responding scales, and performance outcomes. They found that method variance due to socially desirable responding, as measured by socially desirable responding measures, was negatively related to

performance and that it further would suppress personality-performance relationships for self-report measures. They also showed that method variance was partially assessed by socially desirable responding scales, and that the BIDR was a better measure than unidimensional lie scales. However, in addition to this, relative to the effects of self-report method variance, socially desirable responding scales, in general, are also influenced by the Big Five dimensions of Conscientiousness, Emotional Stability, and Agreeableness. This matter will be re-visited in Chapter 6. This provides partial support for the research of Ellingson et al. (2012). Of particular interest is that Connelly and Chang (2016) also state that “The BIDR SDE and IM scales appear to more effectively tap self-report method variance than does the general set of SD scales....Thus, the BIDR’s SDE and IM subscales both appear capable of capturing the effects of self-report response styles on performance” (p. 8).

In discussing the limitations of their research, Connelly and Chang (2016) point out that in high stakes assessments socially desirable responding scales’ ability to assess response styles would be higher than their meta-analysis research showed. That is because, in their view, more variance in response styles would be expected due to the nature of the assessment context. For researchers these findings suggest that SD scales ‘may be salvaged’, to quote Connelly and Chang (2016), if their relation to substantive traits could be reduced. In addition, the Connelly and Chang (2016) findings support those of Lonnqvist et al. (2014) that impression management, as measured by the BIDR-IM scale, is an interpersonal deliberate impression management phenomenon rather than an unconscious self deception phenomenon. Taken together these two studies provide a strong counterargument to the views of Uziel (2014) on the construct validity of the BIDR-IM, a bespoke measure of which was used to detect faking good in the high stakes employment selection context of this

research programme. The major advantage of the research carried out by Connelly and Chang (2016) in assessing the construct validity of the IM measure used in the field study can be seen when the reader recalls the earlier section in Chapter 2 on the methodological advantage of using Campbell and Fiske's (1959) MTMM approach for separating method variance and substantive variance due to traits.

In summary, there are some legitimate concerns about the construct validity of the BIDR arising from extant research. This is due to the overt nature of the items in the measure, and the potential for faking good to occur when responding to the items. On the other hand, there is evidence that the BIDR does capture the occurrence of faking good (Ellingson et al. 2012; Fan et al., 2012). Nevertheless, even if the procedure of using a formal warning in assessing personality in high stakes selection contexts is followed it will still be necessary to measure, by some means, whether faking good has been fully eliminated or just minimised. This is so because the consequential aspect of Messick's (1995) approach to comprehensively establishing construct validity is a necessary requirement. Failure to do so can result in unfair rank order selection effects when individual candidates are selected from a pool of job applicants (Fan et al., 2012). The next section of the chapter examines this issue in some detail.

5.2 Rank Order Selection Effects

Rank order selection effects pose another challenge to construct validity. If faking good occurs with some participants in spite of the effectiveness of the formal warning, and if it can be measured, the construct validity of the NEO PI-R can still be questioned because of the consequential aspect (Messick, 1995). The question of the

external aspects of the construct validity of personality measures, as expanded upon in Chapter 2, is a much broader one than the narrower aspect of criterion related validity (Messick, 1995) because of the consequential effects of unfairness, for instance. The failure to detect faking good in personality assessments in high stakes employee selection situations has consequences (Sackett, 2012). Consistent with the Messick (1995) approach to construct validity Rosse, Stecher, Miller, and Levin (1998) called for personality research to examine the consequential impact of faking on decision making, and not just with respect to criterion-related validity. Even though they found the factor structure of personality to be invariant in their samples the Ellingson, Sackett, and Smith (2001) article points out that there are still legitimate construct validity concerns:

If social desirability is introducing systematic bias into scores, individuals responding in a highly socially desirable manner will obtain artificially inflated scores on various scales. Assuming organizations select individuals from the top down, selection decisions made on the basis of those raw scale scores have the potential to be dramatically influenced. For those scales reporting large effect sizes when comparing the groups, the mean differences would translate into overprediction and the selection of more individuals who are highly socially desirable in their responses, as those individuals will represent the top of the score distributions. If these individuals are intentionally distorting their scores, this implies that organizations are selecting individuals who have not been honest in their responses. (p. 131)

Executive selection is a form of high stakes selection in which the outcome is that of selecting a preferred candidate from a short list or subset of candidates. In essence, the

selection decision rests on a rank ordering, formal or otherwise, of the short-listed candidates. Success criteria for these types of positions can be difficult to define (Highhouse, 1998; 2002). This can make the task of deciding the basis on which to make the selection decision somewhat dependent on intuition and subjectivity (Highhouse, 2008; Hollenbeck, 2009). Highhouse (1998) suggests that a realistic approach to use when it comes to selecting a candidate from a short list of finalists is to use a combination of a formulaic approach together with intuition. It is stating the obvious that if personality measures are used in the formulaic approach, and faking good occurs, this can easily affect the rank order of individual candidates. If the selection decision is based, to some extent, on a particular criterion that is used as a substitute for some actual measure of job performance then an unfair or biased selection decision is likely to occur. The substitute criterion could be, say, the Big Five personality dimension of Conscientiousness, or some composite score that combines a number of personality dimensions, or a linear combination of personality dimensions together with an ability measure such as general cognitive ability. Because the selection criterion or criteria includes personality dimensions then the rank order of candidates can easily change when even one of the candidates deliberately engage in faking good compared to a ranking based on all of the candidates' true scores on these dimensions (Komar, Brown, Komar, & Robie, 2008; Rosse et al., 1998).

Therefore, if using a formulaic approach to making the selection decision when top down selection is used to make the decision, the candidate that fakes good is at an unfair advantage in the selection process and the selection process is both unfair and biased. This situation is worsened if, in addition, cut-off hurdle scores are used in the selection process. So, for example, for conscientiousness a candidate might get eliminated from the selection process due to his/her true score being below the mean score of the norm group candidates on this personality dimension (cut-off point). At the same time, a

different candidate with a true score below the mean but whose reported score is above the mean might remain in contention as a finalist (Donovan, Dwight, & Schneider, 2014; Griffith et al. 2007; Peterson, Griffith, & Converse, 2009). Hough and Oswald (2008) state that “To the extent that effective criterion measurement is not in place, and to the extent we cannot determine the type of test score faking that would lead to harming the organisation (not hiring more effective individuals) and to the qualified applicant (being displaced by applicants who fake more), the effect of faking on criterion related validity becomes a more difficult question to answer. We face these challenges in today’s research” (p. 283).

The consequences of a failure to achieve this objective in high stakes employee selection contexts can be severe. A disproportionate likelihood of those who fake good being selected, which can easily arise from unfairness in testing, has been found consistently in research examining the effect of faking on the rank-order of those selected (Ellingson et al., 2001; Hough, 1998; Komar et al., 2008). This is because, unless dealt with, faking good consistently affects the rank order of applicants such that those who faked are selected at a higher rate than those who were honest in their responding. The evidence for this is reviewed in the next subsection.

5.2.1 Empirical Evidence for a Rank Order Effect

Rosse et al. (1998) found that a disproportionate number of the highest ranking applicants in their research were faking their responses in the personality assessment. With a low selection ratio of .05, as many as 88% of the new hires in their sample may have had significantly lower true Conscientiousness scores than those

reported in the self assessment. Mueller-Hanson, Heggstad, and Thornton (2003) studied this rank order effect with personality measures in a laboratory setting. Participants in the faking group were told that high scorers on the assessment would be selected to participate in the second part of the study in which a cash prize would be given to the top scorers, and also told participants which traits were being assessed before the test was administered. An honest response condition was used as a control group. Participants in the honest group were told that they were completing the assessment for research purposes only. The faking group scored higher in the assessment than those in the honest condition. When they looked at the effect of selection ratios, they also found that as the selection ratio decreased i.e., fewer participants were selected, the proportion of participants selected from the faking group increased. Christiansan, Goffin, Johnston, and Rothstein (1994) studied the effect on rank ordering of supervisory participants completing a personality assessment for future selection, developmental, and other purposes, and found that after correcting personality scores for response distortion, the rank-order changed for over 85% of candidates.

The consequences of a selection decision based on a personality assessment in which a candidate has faked good forms part of the nomological net underlying Messick's (1995) concept of construct validity. The 'social consequences' of the inferences made in assessments are part of this nomological net. It follows from this rank order effect that the impact on behaviour of factors such as sub clinical narcissism and psychopathy, which are related to the Big Five personality traits (Markon, Krueger, & Watson, 2005; Paulhus & Williams, 2002), need to be considered. Both of the sub clinical syndromes mentioned can lead to poor outcomes in organisations (Boddy, 2005; Chatterjee, & Hambrick, 2007; Grijalva, Harms,

Newman, Gaddis, & Fraley, 2015; Jonason, Slomski, & Partyka, 2012; Stevens, Deuling, & Armenakis, 2012). As Chatterjee and Hambrick (2007) make clear narcissism in CEO's is positively related to strategic dynamism and grandiosity, and it engenders extreme and volatile organisational performance. So this is not some trivial artefact of psychometric assessment. The CEO of an organisation determines strategy, and strategy determines future outcomes (Boddy, 2011; Lease, 2006; Padilla, Hogan, & Kaiser, 2007; Resick, Whitman, Weingarden, & Hiller, 2009; Singh, 2008; Stein, 2003). Hogan, Hogan and Kaiser (2011) estimate that the average base rate for managerial failure is 50%. Executives at top levels of organisations are also failing earlier after being promoted, with reported failure rates in the first year and half in position ranging from 16-40% (Zaccaro, Gulick, & Khare, 2008). So employee selection procedures, in general, are arguably not meeting one of Huselid's (1995) requirements for high performance work practices, namely, comprehensive employee recruitment and selection procedures. Hence the practical need for better selection procedures including those relying on personality assessments.

Because the field studies of the research programme of this thesis did not address the issue of consequences of faking on the rank order of job candidates after the procedural use of a formal warning it was necessary to devise a methodology that would capture this component. To deal with this aspect of establishing the construct validity of the NEO PI-R a Monte Carlo simulation approach was used in the research programme in order to address the issue of rank order effects arising from the incidence of faking good, if any, among those participants who nevertheless faked good in spite of the warning.

The approach of Einhorn and Hogarth (1975) in addressing the question of effective criterion measurement where criterion measures are not available is of value

when it comes to decisions in high stakes executive selection. This approach was used in the Monte Carlo simulations where the criteria for determining accurately job performance are not available. They recommend that unit weighting schemes (Camerer, 1981; Dawes, 1979; Wainer, 1976) be used for predictors of job performance given that the sign of the zero order correlations is known.

Even though as Einhorn and Hogarth (1975) point out that “The company officials all have different and even vague definitions of what job success is (a most usual occurrence!) the equal weighting rule will suffice for decision making purposes” (p. 183). However, this approach is effective *only* if the predictors used in the weighting scheme, which constitute the composite criterion, are construct valid. This is another reason why it is vitally important to deal with the faking good issue when it comes to the use of personality assessment in determining the rank order of job candidates particularly for middle and senior management positions. This arguably may be achieved, according to the body of research reviewed for this thesis, by using a construct valid impression management measure in combination with a warning about the consequences of faking good if detected so as to ensure, as far as possible, that only those candidates whose observed scores on the Big Five dimensions of personality are as accurate a reflection as possible of the underlying latent traits are considered for selection.

To summarise, this chapter following on from the research review of the previous chapter explored in some depth conceptual and theoretical aspects of the construct validity of self-report impression management measures. It also examined the evidence for and against using the bespoke version of the BIDR-IM self-report measure for detecting faking good, and why it was necessary from a strong construct valid perspective to use such a measure in the research programme. The next chapter

explores in depth many of the measurement issues that arose from the theoretical reviews of Chapters 2 to 5. It explains in detail how the main methodological issues that arose during the research programme were dealt with both procedurally in the pretesting preparatory phase and as they arose.

Chapter 6

Methodology Issues

The previous chapters set out in detail the broad conceptual and theoretical background, for a number of relevant topics, to the actual field study research programme of the thesis. This current chapter is central to providing an answer to the fourth of the key issues set out in Chapter 1. Its primary objective is to provide an understanding of the statistical and analytical approach followed in the research programme used to evaluate the central idea underling the hypotheses tested, based on the concepts and theory explored in Chapters 2 to 5, i.e. that the use of a formal warning at least minimised faking good and, also, to show that the bespoke impression management measure used is construct valid as defined in Chapter 2.

There were a number of methodological issues, which form part of the nomological net of the constructs measured in this research programme, that had to be dealt with and which this chapter examines in turn. The first one of interest, covered in Section 6.1, was the question of how to separate common method variance (CMV) from substantive trait and, also, error variance. The mathematics of how to do this are well established and are based on Campbell and Fiske's (1959) multitrait-multimethod (MTMM) approach to the issue (Chang, Connelly, & Geeza, 2012). Following this aspect of the mathematics of factor analysis are explored in Section 6.2, which are pertinent to a soundly based understanding of the use of the technique. This was an essential prerequisite for the interpretation of the results of the confirmatory factor analyses (CFA's) carried out (Brown, 2006) in Chapter 8, the

Results chapter. Next, in Section 6.3, the issue of how to methodologically establish the construct validity of bespoke impression management measure used is expanded upon in detail. Then, in Section 6.4, the methods used in the Monte Carlo simulations are described. In this section, the approach used to explore the consequences (Messick, 1995, Embretson, 2007) of selection decisions based on NEO PI-R personality assessments is further explained and expanded upon. Finally, in Section 6.5 the hypotheses tested in the research programme are described.

Essentially the methodological approach taken for this research programme was to rely on some form of mathematical modelling (Rodgers, 2010) wherever possible. It was felt that this was the best method with which to establish the construct validity, internal and external (Embretson, 2007), of both the NEO PI-R and the bespoke impression management measure used. This approach was necessary particularly if the strong approach to the substantive, structural, and generalisability aspects of Messick's (1995) criteria for establishing construct validity is to be successful. If the Big Five dimensions can be shown to share little or no substantive trait variance then it is unlikely that there are any higher order factors superordinate to the Big Five (Chang et al, 2012). Equally, if there are two higher order factors, Stability and Plasticity, superordinate to the Big Five that themselves share little or no substantive trait variance then it can also be argued there is no General Factor of Personality (GFP) because the two first order factors are not correlated (DeYoung, 2006). Establishing this property is required in order to investigate the hierarchical structure of the Big Five as measured by the NEO PI-R in the field studies, and the importance of Campbell and Fiske's (1959) MTMM approach in this regard cannot be over-stressed.

Rodgers (2010) uses the following definition that “a mathematical model is a set of assumptions together with implications drawn from them by mathematical reasoning” (p. 1) to describe a mathematical modelling approach to research. This form of mathematical modelling was central in this thesis to attempting to answer the question of whether the NEO PI-R, as well as the bespoke impression management measure, could be shown to have strong (see Chapter 2) construct validity in high stakes employee selection situations. Building and evaluating statistical and mathematical models encourages creativity, according to Rodgers (2010). The primary objective of the research methodology was to try to establish whether the procedural controls – the formal warning and the measurement of faking good - used in the field research to minimise the occurrence of the socially desirable responding arising from faking good or impression management, were effective or not. In essence, this involves separating common method variance (CMV) from substantive trait effects.

6.1 Separating Substantive and Method Effects

As a result of his research Ioannidis (2007), in a very widely cited paper, claims that “Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias” (p. 696). He suggests that the solution for dealing with this ubiquitous problem, in the behavioural and social sciences, is for large studies or meta analyses with minimal bias to be performed on research findings that are considered relatively established, to see how often they are indeed confirmed. The methodologically best way to achieve this, with

respect to personality assessment, is through studies using both an MTMM and CFA approach which also, ideally, use meta-analysis (Chang et al., 2012).

While there are a number of published research studies into the higher order structure of the Big Five, reviewed in subsection 6.1.1.1 below, there is only one that meets all three criteria, including meta-analysis, which is that of Chang et al. (2012). When it comes to the analysis used in the research programme for this thesis these MTMM studies will be main sources relied upon for an understanding of the trait, i.e. free of contamination from CMV, hierarchical structure of the Big Five. This approach is in line with what Ionnidis (2007) recommends as a desirable pre-study research approach. The claimed shared variance between the Big Five in many extant monomethod studies may simply be, to quote Ionnidis, “the net bias that has been involved in the generation of this scientific literature” (p. 700).

It has been pointed out that the extent to which method effects are prominent, a greater proportion of the variance in observed scores is attributable to the method of measurement relative to the intended construct and construct validity is, indeed, compromised (Conway & Lance, 2010; Brannick, Chan, Conway, Lance, & Spector, 2010). In deciding whether or not to deal with the putative consequences of impression management by faking good it is important to mention that Conway and Lance (2010) also points out that, concerning the covariance between measures of two different constructs measured by the same method, the observed score covariance is the true score covariance attenuated by product of the respective loadings on the two different measures. They go on to state that “the widespread belief that common method bias serves to inflate common method correlations as compared to their true-score counterparts is substantially a myth” (p. 327).

Classical test theory explains when researchers should be concerned about CVM due to socially desirable responding such as deliberate impression management or faking good (Conway & Lance, 2010). To understand the different effects of CMV on measurement Conway and Lance (2010) explain the phenomenon in mathematical terms. They show that in the case in which two latent constructs, X and Y, are measured by the same method M, the observed correlation (r_{XY}) between X and Y can be represented as:

$$r_{XY} = \lambda_{XTx}\lambda_{YTy} \rho_{TxTy} + \lambda_{XM}\lambda_{YM} \dots\dots\dots (E 1)$$

where, TX and TY are the true scores of the latent constructs X and Y. λ_{XTx} and λ_{YTy} are X's and Y's reliability indexes, respectively. ρ_{TxTy} represents the X–Y true score correlation, and λ_{XM} and λ_{YM} represent the effect of the common method M on X and Y, respectively. The λ 's in the equation are standardised with unit variance and represent factor loadings or standardised regression weights. If ρ_{TxTy} is zero the two latent constructs are truly independent of, or orthogonal to, each other (Conway & Lance, 2010). This is the fundamental statistical and mathematical logic of the test of the central hypothesis of this research programme.

If it can be shown that 1) the higher order latent constructs of Stability and Plasticity do exist, 2) they are found to be without any, or little, CMV contamination in the field study of this research, and 3) are not correlated, then it can be inferred that the formal verbal warning, given to participants before they completed the NEO PI-R, was effective in minimising or eliminating faking good. However, it should be borne in mind that it is also possible that while the first term of the right hand side of the

equation, the product of λ_{XTX} and λ_{YTY} , reduces or attenuates the effect of the true correlation between X and Y, ρ_{TXTY} , on the observed correlation r_{XY} it is also true that this reduction may be balanced out by the inflation in r_{XY} that is due to the second term $\lambda_{XM}\lambda_{YM}$ in the equation above (Conway and Lance, 2010). Conway and Lance (2010) provide examples of where this is the case and they maintain that this is what generally happens in research using self-report measures.

Similar dual source effects can be seen to occur with the Campbell and Fiske (1959) MTMM approach to establishing convergent and discriminant validity for a latent construct, as well as the effect of CMV on the measurement of traits. The following formula from Lance, Dawson, Bricklebach and Hoffman (2010) is used to explain these effects in the straightforward case of two traits assessed using two different methods:

$$r_{T_1M_1T_2M_2} = \lambda_{T_1M_1}\lambda_{T_2M_2} \rho_{T_1M_1T_2M_2} + \lambda_{M_1}\lambda_{M_2} \rho_{M_1M_2} \dots\dots\dots (E\ 2)$$

where, T_1M_1 and T_2M_2 are the true scores of the latent construct 1 using Method 1 and latent construct 2 using Method 2, $\lambda_{T_1M_1}$ and $\lambda_{T_2M_2}$ are T_1M_1 's and T_2M_2 's reliability indices, respectively. $\rho_{T_1M_1T_2M_2}$ represents the T_1M_1 – T_2M_2 true score correlation, and λ_{M_1} and λ_{M_2} represent the effect of the method M_1 on T_1 and method M_2 on T_2 , respectively. $\rho_{M_1M_2}$ represents the true correlation between Method 1 and Method 2.

The implications of this formula are that, something which is not mentioned by Campbell and Fiske (1959) in their original paper, the convergent validities arrived at by an MTMM analysis reflect the influence not only of the common trait but also

potentially reflect the influence of correlated methods because of the last term in the right hand side of the equation, according to Lance et al. (2010). If there are different method factors and they correlate positively, then covariance inflation may incur even in multitrait-multimethod analyses. A researcher using an MTMM analysis could still find that the sample nevertheless shows a correlation between X and Y due to method effects arising from socially desirable responding from different sources of method variance, such as for example a difference between halo effect CMV in peer reports and faking good in self-reports, because of the inflationary effect of the second term on the left hand side of the equation above, $\lambda_{XM}\lambda_{YM}$. λ_{XM} is the loading of latent construct X on the common method factor M. The importance of this is that even if a methodologically sound MTMM study finds a small correlation between Stability and Plasticity this does not preclude that possibility that the two latent constructs are not correlated.

When analysing data in monomethod studies, where there is a justifiable concern about the distorting effect of method variance, there are a number of different procedural and statistical techniques that can be used to control for CMV and method biases (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; Malhotra, Kim, & Patil, 2006). One of the techniques, which Podsakoff et al. (2003) looked at, involves controlling for the effects of an unmeasured latent methods factor in a confirmatory factor analysis (CFA). Indicators of latent constructs are allowed to load on their theoretical constructs in a CFA, as well as on a latent common methods variance factor, and the significance of the structural parameters is examined both with and without the latent CMV factor in the model. In this way, the variance of the responses to a specific measure is partitioned into three components: (a) trait, (b) method, and (c) random error. This approach was used in the analysis of the results of the field

studies of this research programme. With the contents of this section of the chapter in mind the next subsection reviews the various multitrait-multimethod (MTMM) studies of the higher order structure of the Big Five.

6.1.1 Review of Extant MTMM Studies on the Higher Order Structure of Personality

The classical MTMM design in combination with CFA and structural equation modelling allows one to quantify the relative influence of method effects and traits on personality measures as well as to test specific models of the structure of latent personality constructs adjusted for method distortions (Connelly & Chang, 2016; DeYoung, 2015). As already mentioned in Chapter 3, and in the previous section, the techniques of CFA, meta-analysis, and MTMM have been used by a number of researchers in looking at the higher order factor structure of the Big Five dimensions of personality (Anusic et al., 2009; Biesanz & West, 2004; Chang et al. 2012; Danay & Ziegler, 2011; DeYoung, 2006; Gnambs, 2013; Reiman & Kandler, 2010; Ruston, Bons, Ando, Hur, Irwing, Vernon, & Barbaranelli 2009; van der Linden, Vreeke, & Muris, 2013). The findings of these studies were relied upon to inform and direct the mathematical modelling of the results of the field studies. This approach was necessary in order to discover whether the procedural controls used to determine if it had been possible to minimise and, possibly, prevent faking good to occur among participants had been successful.

6.1.1.1 Extant MTMM Studies

In a set of studies Biesanz and West (2004) examined the higher order factor structure of the Big Five. The first study used a monomethod approach but tested the

participants at three different times. They also used a multi-occasion method approach to cater for the possibility that participants' self-reports might be affected by transitory factors such as mood or fatigue. The pattern of relationships between the Big Five showed a high, but not absolute level of discriminant validity similar to that found earlier by Digman (1997). The pattern of moderate relationships found between the Big Five dimensions was suggestive of the possibility that one or more second-order factors could provide a more parsimonious account of personality structure. In order to deal with the possibility that the findings of the first study were biased because of CMV they conducted a second study using a MTMM approach. The degree of discriminant validity found in Study 2 between the Big Five traits across different informant types showed that the Big Five traits were not significantly related. Biesanz and West (2004) concluded that monomethod studies are contaminated by CMV and that when this is procedurally catered for by using an MTMM approach there is no evidence for any higher order factors beyond the Big Five. Their research concluded with the statement that "These results suggest that within-informant-type influences (e.g., self-presentation; halo effects) may be largely responsible for the correlations observed between the Big Five traits" (p. 852).

A possible problem with the Biesanz and West (2004) study was pointed out by DeYoung (2006) in that the level of inter-rater agreement in their sample was quite low by comparison with other studies which compared the level of inter-rater agreement. This would decrease the likelihood that significant correlations would be evident among the Big Five in latent space. DeYoung (2006) used two personality measures in his study - a single adjective rating instrument (the Mini-Markers) as well as the short BFI (See Chapter 3) – to test the hypothesis that use of a single-adjective personality measure, such as that used by Biesanz and West (2004), is associated with

lower inter-rater agreement. DeYoung's (2006) MTMM study showed that the average magnitude of the correlations and the number of significant correlations among latent Big Five variables were significantly greater for the BFI than for the Mini-Markers. From this he concluded that this was due to the greater inter-rater agreement associated with the BFI. This finding would explain the failure of the Biesanz and West (2004) MTMM study to find significant correlations among the Big Five, and therefore no higher order factors, according to DeYoung (2006). DeYoung's (2006) research found that a hierarchical model with two uncorrelated latent factors - Stability and Plasticity - above the Big Five fitted the data very well for the BFI MTMM analyses. For single informant personality assessments i.e. a monomethod approach, he found that the two higher order factors were fairly strongly correlated, as other monomethod published studies had, and have, found. Both the DeYoung (2006) and the Biesanz and West (2004) studies show that correlations between Stability scales (N, A, and C) and Plasticity scales (E and O) are only present in ratings by a single rater and fail to demonstrate convergent validity across raters.

Building on these two studies Anusic et al. (2009) developed and tested what they named the HAB model for evaluating MTMM studies of the higher order structure of the Big Five. The model identifies method variance due to halo effects in Big Five ratings. 'H' in the model refers to 'halo' error which the authors define as a disposition to attribute socially desirable characteristics to oneself or to somebody else. The A and B refer to the two higher order factors. They showed that halo bias in self-ratings is a reliable and stable bias in individuals' perceptions of their own attributes, which impacts on the findings of monomethod studies of the higher order structure of personality. They re-analysed the Biesanz and West (2004) MTMM data set using the HAB model and found evidence for one, Plasticity, of the two higher

order factors. They did not find evidence for Stability because, in their words, the design of the study, “lacks statistical power to provide strong evidence that alpha is not a valid personality factor” (p. 1147). This result is similar to what McCrae, Yamagata, Jang, Riemann, Ando, Ono, and Spinath (2008) found when they allowed that “there is some evidence in our studies that a true β exists” (p.543). They also re-analysed the DeYoung (2006) MTMM data sets and confirmed the findings of that study. Kandler, Riemann, Spinath, and Angleitner (2010), as part of a study of the genetic and environmental contribution to method effects in twin studies, used self- and peer report data to examine the higher-order structure of the NEO-PI-R. Their findings were similar to those of Anusic et al. (2009) in that they found no evidence in support of a GFP, and that Stability was only a ‘weak’ higher order factor.

In another article, Rushton et al. (2009) published the results of an MTMM study using self, teacher, and parent ratings for the 65 item BFQ-C, a Big Five measure. Two models were reported on, one having the Big Five loading on a GFP and the other a third order model with Stability and Plasticity loading on a GFP. Both models showed very good fit criteria with the third order model being slightly better than the second order model. The published background information reported on by Rushton et al. was sketchy and other than the comment, “the factor loadings, which ranged from .31 to .80, although substantial error variance was also detected” no details were provided regarding the CFA methodology or modifications to the CFA models tested.

All of the other MTMM studies discussed in this section have detailed the various problems encountered in achieving admissible CFA solutions, which are to be expected when using CFA in MTMM studies (Kline, 2011). The comments that Rushton et al. (2009) made to explain why their findings differed from other MTMM

studies concerned the methodology used and was limited to the comment that “the possibility is that the sample of teenagers studied here was much better known to their raters in the school context of evaluation (by peers, parents, and teachers) than is typically the case” (p. 359). They simply ignored the methodological fact that the hierarchical model they tested had only the two first order factors of Stability and Plasticity loading on a GFP contrary to what is suggested for a CFA hierarchical model identification (Kline, 2011, p.249). The van der Linden et al. (2013) study did use an MTMM approach but relied on an exploratory factor analysis using Principal Component Analysis for the extraction of a putative GFP, and did not attempt to take account of any method variance in the extraction of the GFP from the samples.

In their MTMM, using the NEO PI-R as the personality measure, study Danay and Ziegler (2011) used a slightly different approach. They treated the bias in self ratings of personality separately to that of peer ratings. They argued that the two types of rating are ‘structurally’ different. Peer ratings are interchangeable because a group of friends knows a person in similar situations. Self ratings, on the other hand, are based on a complete sample of situations. They found evidence for a GFP when using a monomethod approach for both self and peer ratings of personality traits. However, there was no evidence of a GFP when an MTMM analysis was carried out, but the analysis did show support for the higher order factors, Stability and Plasticity. They made the case that socially desirable responding due to impression management is the explanation for the finding of a GFP in their monomethod studies. They also put forward the view that the variance in support of a GFP in monomethod studies may be a skill set best referred to as successful impression management, which is why a GFP can be found in monomethod studies whether the assessment is by self-report or peer report. A GFP is heavily dependent of variance-covariance matrix that is analysed

(Brown, 2006, Kline, 2010). This can be context dependent, as in the case of high stakes employee selection situations such as in the research programme for this thesis.

The MTMM studies reviewed above have used either multiple raters or multiple occasions as methods used different Big Five measures, and different samples such as adolescent, adult or twin samples. Chang, Connelly, and Geeza (2012) argue that the models used in MTMM CFA studies are demanding in terms of the number of parameters that are necessary and which have to be estimated. As a result, they are very subject to sampling error and this error can ‘snowball’ across the correlation, or variance/covariances, matrices used in the primary MTMM studies. To overcome this problem they carried out a meta-analytic MTMM study of the Big Five. They tested a number of different models including a GFP model having the Big Five load on a higher order GFP, and a model with the Big Five loading on Stability and Plasticity. The GFP model yielded very poor fit statistics and the authors concluded that “these results indicate that a single general factor beyond the Big Five traits does not effectively account for the covariance between the Big Five trait factors” (p. 10). However, they found that the model including Stability and Plasticity did, “account for the modest correlations that were found between the Big Five trait factors”. The modest correlation that they found between Stability and Plasticity was .12 which is much smaller than the correlations reported in previous monomethod research e.g. from .45 to .72 (Backstrom, Björklund, & Larsson, 2009; DeYoung 2006). Recalling equation E1 on p. 135 this low correlation is not unexpected because as Danay and Ziegler (2011) point out there are method differences between self and peer ratings. Podsakoff et al. (2003, p. 880) also showed that even if the true correlation of two constructs was zero the observed correlation could easily be greater than zero solely because of CMV. Peer ratings are more subject to halo effects, whereas self ratings

are more subject to other method effects as well as halo. Therefore the $\lambda_{M1}\lambda_{M2} \rho_{M1M2}$ term in the Lance et al. (2010) equation on p. 136 would mean that Stability and Plasticity could still be found in a study to be correlated to some degree because of method effects arising from different raters.

The most recent MTMM study on this topic is that of Gnambs (2013). He argued that if there is indeed a GFP that is not simply due to CMV but is a substantive higher order latent construct of personality, it should be unaffected by the length of acquaintance. Failure to find evidence in support of a GFP at long-term acquaintance, even if it is identified at short-term acquaintance, implies that a GFP is more likely to be a product of stereotype based judgments. Gnambs found that a putative GFP could be extracted from ratings of dyads who had known each other for a comparably short period, but that it gradually disappeared with increasing length of acquaintance. On the other hand, he showed that there was evidence that the Stability/Plasticity higher order factors do exist in latent space. The support for Plasticity was strong in both short-term and also long-term acquaintance groups and seemed to be better defined in pairs that knew each other a longer time. It also replicated across cultures. The existence of Stability was also supported in that the loadings for Neuroticism and Conscientiousness gradually increased for long-time acquainted dyads, whereas the respective loading of Agreeableness continually decreased. Stability was clearly identified in North American samples but was ill-defined among European samples.

With the exception of the Ruston et al. (2009) and Van der Linden et al. (2013) studies, none of the other MTMM studies found any support for a GFP but did find some evidence for the putative existence of the two higher order factors, Stability and Plasticity, that were not correlated to a meaningful degree. They also support a putative hypothesis that Stability and Plasticity may not be correlated, or if they are

the correlation is very small and may be due to method factors being correlated (see equation E1 on p. 135 and equation E2 on p. 136). Evidence in support of this latter point comes from the results of a meta-analysis by Connolly, Kavanagh, and Viswesvaran (2007) which showed that both self and observer ratings have a high degree of construct overlap as well as substantial unique variance.

As a result of this review of MTMM studies the strategy that was used in the analyses of the results of the field studies, in Chapter 8, to determine if the formal verbal warning was effective in minimising or eliminating CMV due to faking good was to examine the correlation between Stability and Plasticity. If there was evidence for two higher order factors that were not correlated then it is argued, based on the review of MTMM studies, that it can be inferred that the formal verbal warning was effective in minimising or eliminating faking good on the part of participants in the field studies. In order to see if that was indeed the case in the analyses carried out a good understanding of the techniques of factor analysis was required. This topic is next examined in some detail.

6.2 Factor Analytic Considerations

The second issue in dealing with how to separate and measure trait, method, and error variance is that it is best done using the technique of factor analysis. Initially Campbell and Fiske's (1959) methodology was used in research by relying on a visual inspection of the MTMM correlation matrix. Because of advances in software MTMM analyses can now be based on the application of confirmatory factor analysis (CFA) without too much difficulty (Byrne, 2010; Chang et al., 2012). Factor analysis has played a fundamental role in the development of measures that are used to assess

personality, as was pointed out in Chapter 3. It was Cattell's factor analytic work in the 1940's that led to the first major omnibus personality inventory, the 16PF (Digman, 2002). Initially personality research relied on exploratory factor analysis (EFA), which is inductive and data driven. In more recent times CFA has played a major role in test construction and, more recently, as a diagnostic tool in structural equation modelling (SEM). According to Vassend and Skrondal (2011), the main role of EFA is to contribute to the generation of theories regarding the measurement of latent variables whereas CFA is required for the empirical testing of such theories.

CFA, unlike EFA, is deductive and is used to confirm a priori hypotheses based on theory. It relies on the fundamental principle of local independence i.e. the idea that manifest variables are unrelated to each other when controlling for the common factor. This means that if F is a latent trait then the manifest responses to the items measuring that trait are independent in a subpopulation in which F is fixed. CFA techniques estimate the parameters of a model based on the covariances and variances of the data. If the parameters - factor loadings and residual variances (Brown, 2006) - have a unique identity they are said to be identified. If there is more than one common factor, as in the case of Stability and Plasticity the higher order Big Five factors, then they will only be identified using CFA (McDonald, 1999; McDonald & Ho, 2002) under the following general conditions:

1. For each factor there are at least three items or tests, with nonzero loadings, that have zero loadings on all other factors.
2. For each factor there are at least two items or tests, with nonzero loadings on all other factors, and also, any factor having only two defining items or tests is correlated with other factors.

According to McDonald, “these conditions are possibly not as widely known as they should be by researchers using factor analysis *The conditions are very likely to be satisfied in careful test construction, as opposed to the exploration of the structure of general psychological attributes*” (italics added) (p. 179). The issue is very pertinent to the evaluation, in this programme of research, of the relationship found between the Big Five dimensions of personality and putative higher order factors using CFA. As seen from the review in Chapter 3, Stability is putatively determined by the covariances of three of the five – Conscientiousness, Agreeableness, and Neuroticism. Plasticity is putatively determined by the covariances between Extraversion and Openness (DeYoung, 2007). If there is no GFP at the apex of the factor hierarchy i.e. Stability and Plasticity are not correlated then, according to McDonald (1999), local independence may be an issue at the level of these two uncorrelated factors because of condition 2) above. However, if Stability and Plasticity are correlated then condition 2) above is met, which implies that because there may be a GFP because the lower order latent metatraits are correlated.

Furthermore Kline (2011), in reference to ‘non standard’ CFA models i.e. those where some indicators crossload on more than a single factor or some error terms covary, draws attention to the rules of Kenny, Kashy and Bolger (1998) for nonstandard confirmatory factor analysis models with correlated measurement errors. A CFA model is "identified" if the known information available implies that there is one best value for each parameter (e.g. factor loadings and correlations, indicator uniquenesses) in the model whose value is not known. The parameters of a CFA model are generally considered identified if the researcher can solve the covariance structure equations for the unknown parameters. If more than one solution exists, the

parameter is overidentified. A model is said to be 'just' identified if it is possible to estimate a single, unique estimate for every free parameter. If there is no solution, it is underidentified (Brown, 2006; Kline, 2011). Both just identified and overidentified parameters are labelled 'identified'. Both types of model provide unique values for the parameters, since the multiple solutions for the overidentified parameters can all lead to the same value for the parameters when based on the population covariance matrix of a correctly specified CFA model. If all of the parameters in a model are identified, the model is said to be identified.

In just identified models the numbers of knowns equals the number of unknowns. There are zero degrees of freedom. Such model has a single unique solution in that there is a single set of parameter estimates that perfectly reproduce the input covariance matrix from the sample (Brown, 2006). For overidentified models there are fewer parameters than there are observations, therefore a number of model solutions can be arrived at and compared (Kline, 2011). In 'empirically underidentified' models the conditions for just or over identification are met but it is still not possible to obtain a set of parameter estimates that is both valid and unique (Kline, 2011). The CFA analysis will either fail to find a solution or will arrive at an 'improper' or 'inadmissible' solution containing Heywood cases i.e. negative variance or a correlation greater than one (Brown, 2006; Kline, 2011). Both of these foregoing points will be seen in Chapter 8 to have relevance to the outcome of the research programme.

The Kenny, Kashy and Bolger (1998) rules set out the identification requirements in 'nonstandard' CFA measurement models e.g. those that have correlated errors. Because correlated errors can be due to factors such as the cross loading of factor indicators, which is likely when using the NEO PI-R (Costa & McCrae, 1995), this is

an important matter to consider. There are three conditions, all of which must be satisfied in order to identify non standard CFA models (Kline, 2011):

1. “For each factor, at least one of the following must hold:
 - a. There are at least three indicators whose errors are uncorrelated with each other
 - b. There are at least two indicators whose errors are uncorrelated and either:
 - i. The errors of both indicators are not correlated with the error term of a third indicator for a different factor, or
 - ii. An equality constraint is imposed on the loadings of the two indicators.
2. For every pair of factors there are at least two indicators, one from each factor, whose error terms are uncorrelated.
3. For every indicator there is at least one other indicator (not necessarily of the same factor) with which its error term is not correlated” (p. 140).

Kenny, at http://davidakenny.net/cm/identify_formal.htm, lists other requirements over and above those listed above from Kline (2011) which deal with the one indicator variable case, correlated latent variables, factor loading, and indicators that load on two factors.

As well as issues concerning identification, the question of correlated errors is very pertinent to the application of CFA to omnibus personality inventories because, as Johnson (1994) has shown, there is strong evidence that the facets of both the NEO PI-R and the HPI (Hogan & Hogan, 1995) load on two factors in many cases, and

Markon, Krueger, and Watson's (2005) research on the unbalanced nature of higher order factor structure of the Big Five, so correlated errors are likely to be present in the results of this research programme. This can be easily seen by examining the AB5C factor loadings structure of the facets Extraversion and Conscientiousness for the NEO PI-R (Johnson, 1994). In AB5C terms, a trait's facets are depicted by their loadings on two factors, a primary and a secondary, that best describe it as referred to earlier in Section 3.2 of Chapter 3. The primary factor loading is signified first in the table below, and the secondary loading follows that in the notation as shown below. The + and – negative sign indicates the pole of the dimension on which the facet loads.

Table 2
Primary and secondary factor loadings of the Facets of Extraversion and Conscientiousness

Extraversion		Conscientiousness	
Warmth	E+A+	Competence	C+N+
Gregariousness	E+N+	Order	C+O-
Assertiveness	E+C+	Dutifulness	C+A+
Activity	E+C+	Achievement Striving	C+E+
Excitement Seeking	E+A-	Self Discipline	C+N+
Positive Emotions	E+E+	Deliberation	C+E-

The AB5C model is arguably the 'gold standard' against which to evaluate Big Five measures when looking for theoretical rationales for the existence of correlated errors between indicators of latent higher order factors of personality (Hofstee, De Raad, & Goldberg, 1992; Johnson & Ostendorf, 1993).

Reilly and O'Brien (1996) point out that demonstrating the identification of parameters in measurement models which include correlated error terms presents a serious challenge for researchers using CFA and structural equation models. If a parameter is not identified, then it is not possible to obtain a unique point estimate of

its value. So the evaluation of omnibus Big Five personality measures using CFA is inherently problematic as McCrae, Zonderman, Costa, Bond and Paunonen (1996) pointed out. For example, Kline specifically states that “To identify a hierarchical CFA model, there must be at least three first-order factors” (p. 249). McDonald (1999) in the first sentence of his section on ‘Higher Order and Hierarchical Factors’ states, “Suppose now that we have an independent clusters model with at least three correlated factors” (p. 188), yet as seen earlier Ruston et al. (2009) found evidence from an MTMM CFA for a hierarchical factor structure with two first level factors loading on a second level GFP. In addition, Monte Carlo simulation by Marsh, Hau, Balla, and Grayson (1998) showed that if the number of indicators per factor in a CFA study with more than one factor is increased from a minimum of two per factor there were more proper solutions and more accurate parameter estimates.

Ideally, CFA is used in a confirmatory manner to construct a psychometrically sound psychological measure with a clear (congeneric) factor structure based on items in the measure that are strictly unidimensional (Brown, 2012; McDonald, 1999). This approach to test design does not apply to any of the, already constructed, omnibus personality measures in commercial use (Hopwood, Wright, & Donnellan, 2011) and this consideration can present major methodological problems when arriving at conclusions using CFA to explore questions concerning the construct validity of any of these measures. Therefore without a thorough and informed examination of a wide range of evidence from the nomological net (Cronbach & Meehl, 1955; Embretson, 2007) surrounding a latent construct, it is easy to arrive at erroneous conclusions regarding the construct validity of a measure such as the NEO PI-R in high stakes employee selection situations. However, it is also important to make the point that in spite of this CFA can be used as a mathematical modelling device or diagnostic tool

for testing one explanatory model against another (Rodgers, 2010). This is the approach followed in the analysis of the results in this research programme.

The foregoing review of factor analysis makes it clear that there were many methodological pitfalls to be cognisant of in using CFA for the analysis and interpretation of the results of the NEO PI-R assessments of participants that, as will be seen later in Chapter 8, had to be taken into account in the analysis carried out on the results of the field studies in this research programme. Hence, an understanding of the CFA technique and, in particular, aspects of the mathematical basis of the software used for carrying out a CFA was necessary for arriving at a detailed understanding and analysis of the field studies' results of the research programme. In addition to the factor analytic issues discussed above, there are also important methodological issues concerning the impression management measure used in the research programme, which are reviewed in the next section of the chapter.

6.3 The Construct Validity of the BIDR-IM Scale

Since there have been questions raised by some researchers (Biderman & Nguyen, 2009) about the use of faking good measures this issue needs to be as fully explored as possible in order to make the case that the bespoke BIDR -IM measure used in the research programme was construct valid. The use of an impression management measure played an important role in the analyses of the results of this research programme. Yet MacCann, Ziegler, and Roberts (2012) and others (Morgeson, Campion, Dipboye, Hollenbeck, Murphy & Schmitt, 2007; Uziel, 2014) have recommended against the use of such measures as an indicator of faking good. Because of this, the issue will have to be examined in some depth in order to make the

case for the construct validity of the measure used. The first issue to be considered in that regard is the one of restriction of range.

6.3.1 Restriction of Range Issues

One of the methodological issues which arose in the research for this thesis, affecting construct validity inferences arising from the use of a faking good measure, is that of the statistical phenomenon of restriction of range (Murphy & Davidshoffer, 1997; Hunter, Schmidt, & Le, 2006). Its effect on the determination of the construct validity of the faking good measure used in the research programme is of importance. The reason for this is that the determination of the validity of the inferences arising from the impression management measure used in the research programme, as well as the NEO PI-R, is based on a restricted sample rather than a population based one. Measurement of the extent of participants' faking good was based on the BIDR-IM measure (Paulhus, 1984), using data that was arguably more representative of the general population than was the participants' personality data which clearly came from a restricted population. It will be seen later in this section that this is an important consideration with respect to establishing construct validity.

Whenever a sample has a restricted range of scores the correlation of a latent construct with a criterion of interest will be reduced or attenuated (Murphy & Davidshofer, 1998). The participants in this research programme were all applicants, with relevant work experience and qualifications, for middle and senior management positions in a range of organisations. Their mean scores and standard deviations on the Big Five, which were assessed in the research programme, were different from

that of the population at large – higher mean scores and lower standard deviations - as will be seen in Chapter 8. This is what the ‘gravitational hypothesis’ (Wilk, Desmarais, & Sackett, 1995) would predict with respect to a natural tendency for the environments and psychological attributes of individuals to align to some degree in occupational settings. Connelly and Chang’s (2016) meta-analytic findings and conclusions concerning what the BIDR-IM measures are for an unrestricted population. To arrive at their conclusions they had to make adjustments, which accounted for the restriction of range effect, to the reported correlations of the individual studies included in their meta-analysis, in order to carry out the meta-analysis. A mathematical explanation for the need to make adjustments to the reported studies used in the meta-analysis can be found in Hunter, Schmidt, and Le (2006).

There are two types of restriction of range that can occur – direct and indirect (Salgado, 2016). Direct restriction of range occurs in predictor/criterion relationships between two variables, say x and y , when variable x is from a truncated (restricted population) and variable y is not. The correlation between the two variables r_{xy} is lower in the restricted or truncated population than it is in the unrestricted population (Sackett, Lievens, Berry, & Landers, 2007). In the case of indirect restriction of range the truncation that occurs affects the observed correlation between predictor variables. Indirect restriction in predictor/criterion relationships can have a much greater effect than direct restriction, according to Sackett et al. (2007).

Take the hypothetical case of some outcome criterion of interest that has three predictors, A, B, and C. Assume that each of the three predictors is correlated with the criterion in the population and that A and B are intercorrelated, but are not correlated with predictor C. Both direct and indirect restriction of range occurs in this situation. In employee selection situations truncation i.e. eliminating cases, occurs because of

the use of cut-off scores and/or because of selection from an otherwise restricted rather than unrestricted population. This truncation leads to restricted range effects that are both direct and indirect (Hunter et. al, 2006). As a consequence the correlations that Connelly and Chang (2016) found in their meta-analysis between a number of the Big Five dimensions of personality and BIDR-IM scale scores were impacted on by both direct and indirect restriction of range. Agreeableness and Conscientiousness, as measured by NEO PI-R, have been shown to be intercorrelated (.24) in the general population (Costa & McCrae, 1992). Connelly and Chang (2016) in their meta-analysis found that Agreeableness had an adjusted (for restriction of range) interpredictor correlation of .27 with IM, as measured by the BIDR. Conscientiousness has an adjusted interpredictor correlation of .31 with IM. Their research findings also showed that Agreeableness and Conscientiousness are predictors of BIDR-IM scores when the BIDR-IM is treated as a criterion. The separate latent self-report method factor included in their CFA models, which was not correlated with Agreeableness or Conscientiousness, is also a predictor of the BIDR-IM as a criterion. In addition Connelly and Chang (2016) showed that a structural equation model, containing BIDR-IM scores as a partial mediator, with the Big Five and a latent self-report method factor as predictors of performance, had acceptable model fit. This suggests, by analogy with an A, B, and C example above, that the correlation of Agreeableness and Conscientiousness with the BIDR-IM scale as a criterion *in a particular context* will differ (be lower) compared to the estimate of a population predictor/criterion correlation arrived in the meta-analysis.

This latter outcome is the reverse of what happens in a meta-analysis in which the restricted single study sample correlations are adjusted upwards. In unrestricted populations both direct and indirect restrictions of range effects on the personality

measures are taken into account in adjusting the restricted single study sample correlations upwards to estimate the unrestricted meta-analytic correlations (Hunter et al., 2006; Sackett et al., 2007). High stakes selection contexts, on the other hand, are restricted samples and the correlations found will differ from Connelly and Chang's (2016) unrestricted population estimates. In addition, the latent self-report method factor correlation with the BIDR-IM score will have a smaller direct restriction of range effect than Agreeableness and Conscientiousness in the restricted high stakes situation, and is not affected by an indirect restriction of range effect. Therefore its predictor/criterion correlation in a single study will be less affected in the same context than Agreeableness and Conscientiousness. This is what may have prompted Connelly and Chang (2016) to state that "we would expect SD scales' ability to assess response styles to improve in applicant contexts" (p. 12).

In addition, it can be argued that the use of a pre-test warning may well also have attenuated the interpredictor correlation of Agreeableness and Conscientiousness in the field study. Since the adjustment to the restricted interpredictor correlation arises from the truncation of scores on some of the personality dimensions, and arguably less so from the truncation of 'scores' on the latent self-report method factor of Connelly and Chang (2016), the variance accounted for by these personality dimensions will be reduced in a single study such as this field research, while that due to the self-report method factor may not because of either no, or a lesser, restriction of range effect. Based on the evidence concerning moral hypocrisy and moral disengagement, described earlier in Chapter 4 (Mazar, Amir, & Ariely, 2008; Shu, Mazar, Gino, Ariely, & Bazerman, 2012) it is also arguable that variance due to personality is even further reduced in addition to the restriction of range effect highlighted above. This is because of the impact of the saliency effect, arising from the

formal warning used in the assessment test battery (see Chapter 7), found by Batson, Thompson, Seuferling, Whitney, and Strongman (1997) in their Study 3 research findings, which was described in Chapter 4, Section 4.2.1.

Another way of looking at the question of the construct of the impression management measure is to suppose that a putative criterion ‘faking good’ (I) score can be accurately measured and that it is predicted by a composite personality score (P) as well as a self-report method factor score (F). Connelly and Chang (2016) have shown that the criteria of both job performance and academic performance are affected by an individual’s score on the BIDR-IM scale as well as the individual’s Agreeableness and Conscientiousness scores. They found that the self-report latent method factor (F) accounted for a substantial proportion of the BIDR-IM scale (I) score for any individual, i . This result can be modelled, for the purpose of explanation, by a combination of the true score on P and the true score on F, plus error (McDonald, 1999, p.177) i.e.

$$I_i = \lambda P_i + \gamma F_i + \varepsilon_i \dots\dots\dots (E\ 3)$$

where, λ and γ are factor loadings, and ε is the error term. Since CFA and structural equating modelling are based on an analysis of a variance/covariance matrix anything that reduces the variance on P_i will also increase the proportion of variance in I that is contributed by F (Kline, 2011). Factor scores for F and observed scores for I were found to have convergent validity by Biderman and Nguyen (2009) in their research. Clearly in E3 as P_i approaches zero F_i becomes a better measure of I_i . As mentioned earlier Connelly and Chang (2016) found that including the BIDR-IM scale as a mediator, when examining the criterion related effect on performance in their meta-analysis, resulted in a statistically significant improvement in their SEM model fit

compared with a model in which performance was predicted by personality and a latent self-report method factor. The procedural use of the formal warning together with the use of the bespoke version of the BIDR-IM, described later on in subsection 6.3.3, may have contributed to making this mediator effect larger in the field research of this thesis than Connelly and Chang (2016) found in their meta-analysis.

In the particular case of the field studies of this research programme the context in which the impression management measure was used was specific to the administration of the personality measure after the formal warning was given in the assessment exercise for a restricted range participant sample. As mentioned earlier, arguably from an impression management measurement perspective the sample was a less restricted one. This is because the environmental factors that impact on the Big Five personality dimensions means and standard deviations were not the identical factors to those that may have affected the participants' IM scores, and those that were arguably did not do so to the same extent. Inferences about the construct validity of the BIDR-IM scale as a valid measure of faking good have to be contextualised as Connelly and Chang (2016) pointed out that "it is possible that the combination of substance and style within SD scales may shift across contexts" (p. 12). 'Substance' refers to actual high standing on traits that are desirable, whereas 'style' refers to deliberate dishonest responding (Chang et al., 2012). This substance and style contextual effect is consistent with the research of Ziegler and Buehner (2009) who also found that the faking good effect in their research was related to personality as well as a self-report method factor. The Connelly and Chang (2016) meta-analysis of socially desirable responding and, in particular, of the BIDR showed that the BIDR-IM scale does indeed account for meaningful variance in responding due to a socially desirable responding method effect in the population at large. The procedure under

which the BIDR-IM scale was used and administered in the field studies of this research programme was bespoke with respect to the objective of the assessment exercise and the restricted participant field study sample. This is likely to have reduced the variance in the field study due to personality dimensions that Connelly and Chang's (2016) meta-analysis found to partially account for scores on the BIDR-IM scale in their meta-analysis. That this reduction may have occurred can be inferred from both the Hunter et al. (2006) and the Sackett et al. (2007) articles on how to account for both direct and indirect restriction of range.

The combination of the effect of direct and indirect range restriction with two additional factors help to make the case for the construct validity of the faking good measure used in the research programme. The first factor was the use of the bespoke IM measure used arising due to the embedding of distractor items in the questionnaire used which is covered in subsection 6.3.2, below, of this chapter. The second factor was a contextual effect, explored in subsection 6.3.3, arising from the warning used with the personality assessment in the administration of the battery of tests used. These three factors taken together arguably support the contention that the bespoke BIDR-IM scale as used in the field study has good convergent validity with a putative latent SD construct.

6.3.2 The Item Transparency of the Impression

Management measure

Procedural controls, according to Podsakoff, MacKenzie, and Podsakoff, (2012), are needed to be in place in order to minimise the degree to which the IM

measure is conflated by method variance due to the measure ‘grouping related items together’ (p. 552). The BIDR-IM scale contains 20 items, half of which are reversed scored. The items in the scale are, deliberately, both overt and clear cut. This presents a construct validity concern (Embretson, 2007) in ‘high stakes’ selection situations – presenting the twenty items as a group to subjects in selection situations was a concern simply because of the overt and clear-cut nature of the items (Podsakoff et al. 2003).

The research findings concerning moral hypocrisy of Chapter 3 would suggest that this effect will be manifested in the context of the assessments of participants in this programme of research. Based on the research of McFarland, Ryan and Ellis (2002) into the effect of random item placement compared to item grouping in self-report questionnaires it was felt that participants who were likely to engage in faking good would be likely to recognize the objective of the IM items if they were presented as a single group of 20 items. The psychometric properties of the personality measure used by McFarland et al. (2002) were found to be better when the items that measure the same construct were randomly distributed throughout the test. There was an additional reason for this approach arising from coaching for tests such as the BIDR. Hoffman (2006), in a widely available book published to help test takers, provides a list of sixty-eight ‘softball’ questions similar to those items in the BIDR to help the intending faker spot the items designed to detect faking good.

To deal with these concerns in the research programme, rather than grouping them together the 20 IM items were randomly included in a bespoke questionnaire of 75 items containing the 20 IM items and 55 distractor items. The 55 distractor items were not related to impression management. Rosse, Stecher, Miller, and Levin, (1998) used a similar approach when using the IM scale of the BIDR. This bespoke nature of

the modified BIDR-IM scale arguably helped to minimise a socially desirable responding effect from contaminating the accuracy of the IM measure used in this research programme. It is not clear from the Connelly and Chang (2016) meta-analysis the number if any, of studies used in their analysis, to take this precaution against contamination resulting from socially desirable responding in the BIDR-IM measure itself.

As mentioned earlier, because the BIDR-IM scale is a self-report measure it too would be subject to method variance resulting from socially desirable responding in the same manner as Big Five self-report measures of personality are prone to. The more readily the construct being assessed can be identified from a reading of the items the more likely it is for socially desirable responding to occur (McFarland et al., 2002). McFarland et al. (2002) also found that the psychometric properties of a measure can change across item formats and instruction conditions. The random placement of IM items in the bespoke measure containing the distractor items should help towards achieving the procedural objective of detecting and minimising the effect of faking good. In addition, as Connelly and Chang (2012) point out the context in which job applicants are assessed can impact on the construct validity of the BIDR-IM, as the ‘Latent Process Studies’ element of Embretson’s (2007) construct validation model would suggest. This issue is investigated in the next subsection.

6.3.3 The Context Effect on the Construct Validity of the Bespoke BIDR-IM measure used

The approach taken in this research programme by which the nomological net for the bespoke BIDR-IM scale in a high stakes employee selection context was

established is consistent with what Embretson's (2007) dynamic approach to construct validity requires. It is also in keeping with Cizek's (2012) definition of validity that "Validation is the ongoing process of gathering, summarizing, and evaluating relevant evidence concerning the degree to which that evidence supports the intended meaning of scores yielded by an instrument and inferences about standing on the characteristic it was designed to measure" (p. 140). Context based influences are situational factors that alter test responses by influencing respondents' motivations and goals, or by modifying aspects of respondents' cognitive or emotional processes during testing (Bornstein, 2011). The context based influences of the field studies – high stakes environment, restriction of range, the use of a formal warning – mean that inferences made about the construct validity of the IM measure used in the research have a very important role to play in determining the relationship of the bespoke BIDR-IM measure used in high stakes selection contexts to a putative latent construct of 'faking good'. All of these context based influences affect test performance (Embretson, 2007) and need to be thoroughly understood and accounted for as far as possible. Section 6.3.1 only considered the restriction of range influence.

The previous paragraph is a good example, from a dynamic construct validity perspective, of a debate that sometimes surfaces in the literature concerning the existence and stability of latent traits across situations, such as the situations that Connelly and Chang's (2016) meta-analytic findings can be applied to. An extreme situationalist view, for example, as advocated by Mischel's (1969) claims, with respect to personality traits, that manifestations of traits are solely situation specific. In this view, traits do not exist except as manifestations of situations, and any observed stability of traits is solely a consequence of stability of situations. A well reasoned refutation of this extreme position of Mischel was provided by Roberts and

Caspi (2001) which showed that a lack of understanding of the fact that even though mean scores on a measured variable can change, some dramatically, across situations it does not follow that the rank order of individuals change across situations of interest. Borghans, Duckworth, Heckman, and ter Weel (2008) present a framework that explains this which is equally applicable to personality inventories, IQ test scores, or other psychological measures. It sheds further light on the reason why the bespoke BIDR-IM scale might well show high convergent validity with a putative ‘true’ measure of lying by job candidates in high stakes employee selection situations. Given the important role that the IM measure played in the methodology followed in the field research and subsequent analysis a full understanding, from a mathematical modelling perspective, is warranted (Rodgers, 2010).

The mathematical exposition which follows is based almost entirely on the Borghans et al. (2008, p. 991-992) explanation of why situational effects can be of importance when making inferences about latent constructs. In their notation, f is a vector of latent traits and f_l is a particular trait in the list of L traits (one of the Big Five, for example). The manifestation of trait l , M^n_l , as opposed to the trait itself, f_l , is obtained by measurement n , $n = 1, \dots, N_l$, and may depend on environmental or contextual incentives to manifest the trait.

If R^n_l represents the reward for manifesting the trait l in situation n then it follows from this that if one of the Big Five is a desirable trait in n , and is also highly rewarded, then there will be more manifest evidence of that trait in n , compared to less highly rewarded situations. Reward can be interpreted very broadly to include environmental factors such as the benefits of social approval, the approval of external observers, and/or economic reward – all of which apply in high stakes selection situations. Other latent traits besides l may affect the manifestation of a trait for l . For

instance, a person who has a higher score on a cognitive ability measure may perceive the benefits of exhibiting her or his high level of Conscientiousness in situation n . Let f_{-l} (signifying not f_l) be the components of f that are not f_l . Let W_l^n denote other variables operating in situation n that affect measured performance for l . Observed traits are imperfect indicators for the latent traits that they have been shown to measure, because of measurement error which can vary depending on f_{-l} , R_l^n , W_l^n :

$$M_l^n = h_l(f_l, f_{-l}, R_l^n, W_l^n), n = 1, \dots, N_L, l = 1, \dots, L.$$

This equation shows that individuals in different roles, and with different incentives, manifest differences in observed behaviour. It captures the effects on measurements of the level of the trait (f_l), the incentives (R_l^n) in a situation and the context (W_l^n). It is consistent with Embretson's (2007) dynamic model of the construct validation process. It can be applied to both the assessment of personality in high stakes selection situations and the assessment of faking good using the bespoke BIDR-IM scale in the same situation in the field studies of this research programme. As well there may be threshold effects in all variables, such as floor and ceiling effects (Hauenstein, Bradley, O'Shea, Shah, & Magill, 2017), so the function h_l allows for jumps in manifest traits as the arguments of the equation above are varied. M_l^n will also vary across individuals because of individual differences. Mischel's (1969) very questionable claim that h does not depend on f_l because there is no f_l (or, for that matter, f_{-l}) and, indeed, that the manifestation M_l^n is solely a function of situational incentives, R_l^n , and context, W_l^n , is incorrect. It is a more reasonable argument that, given f_l , the stability of measured traits is a consequence of stability of incentives and context. According to Borghans et al. (2008), the equation above, in

the general case, shows clearly that it is dangerous, from a construct validity perspective, to equate the measurement f_l of a latent trait with the trait itself without standardising incentives and context. It is only meaningful to define measurements on f_l at benchmark levels of R_l^n , \underline{f}_l , and W_l^n . If these benchmarks are defined as \underline{R}_l^n , \underline{f}_l , and \underline{W}_l^n , respectively, then

$$M_l^n = f_l, \text{ for } R_l^n = \underline{R}_l, f_l = \underline{f}_l, \underline{f}_l = \underline{f}_l, W_l^n = \underline{W}_l, n = 1, \dots, N_l, l = 1, \dots, L$$

This is a mathematical operational definition of latent traits across measurement situations. Sjöberg (2015) recently provided some empirical evidence for the validity of Borghans et al.'s (2008) mathematical formulation. He found that the degree of faking that occurred depended on where the test conditions stood on the spectrum of stakes, which ranged from low to high.

As an example of the ceiling and floor effect in the Fischbacher and Heusi (2013) experiments with dice, described earlier in Chapter 4, 20% of participants lie to the fullest extent possible while 39% of subjects were fully honest. In the Pruckner and Sausgruber (2013) field study 39% of customers paid nothing for their newspaper, 42% made a payment that was below the price of the paper, and 19% paid the full price. There was both a ceiling and floor effect in both of these experiments which is consistent with what the \underline{R}_l term, the benchmark level, in Borghans et al.'s (2008) formulation above suggests. There was a benchmark level above or below which respectively the manifest behaviours, such as either not paying anything or paying the cover price, were exhibited. Above or below the benchmark level the price paid was a categorical variable rather than a continuous one even though individuals taking or buying the newspaper would have exhibited a continuous distribution with

respect to individual differences in personality traits. Therefore individual differences in the Big Five dimensions, for those who either paid nothing or paid the full price, didn't matter once the tipping point was reached. From the perspective of the field studies of this research programme, this concept supports the use of cut-off scores, when the behaviour involves some element of moral hypocrisy, in the analyses carried out and described later on in Chapter 8.

The Borghans et al. (2008) mathematical model accounts for the diversity of measurement outcomes that can be encountered in research and applied settings for the same latent trait but in different settings, such as that seen in studies in the Connelly and Chang (2016) meta-analysis of the BIDR measure. It is flexible enough to capture interactions among the traits and the notion that at high enough levels of certain traits, incentives (R^n) might not matter whereas at lower levels they might. It also helps to vindicate the use of the bespoke BIDR-IM scale in the field study for this thesis. Thus, if the trait in question is faking good scores on this latent trait of faking good might also depend on the levels of Conscientiousness and Agreeableness to a, context dependent, greater or lesser extent of the test taker, as Connelly and Chang (2016) have shown. For example, individuals with higher levels of Conscientiousness may engage in faking good, whereas those with lower levels of Conscientiousness may not. According to Borghans et al. (2008), psychologists have not always been careful in characterising the benchmark states at which standard measurements are taken, such as those that form the basis of the Connelly and Chang (2016) meta-analytic findings. This can substantially affect the transportability of tests to other environments beyond that of the test-taking environment (Embretson, 2007). Persons responding to items in an impression management measure in a non high stakes testing environment that did not include a formal warning have different incentives to

respond compared with the participants in the field study, who are being considered for a job at middle or senior management level.

The foregoing detailed exposition of the Borghans et al. (2008) mathematical model was used to provide additional support for the argument that, for the reasons described earlier in this chapter, the bespoke version of the IM scale of Paulhus's BIDR is an acceptable measure from a construct validity perspective, of the putative latent construct of socially desirable 'faking good' method factor. Therefore in the analyses of the results of the field studies of this research programme the bespoke BIDR-IM scale was used to differentiate those who faked good from those who didn't. As well as the issues covered in Section 6.3 reference has already been made on a number of occasions in this thesis to Messick's (1995) inclusion of the consequences aspect of construct validity. Monte Carlo simulation was used to investigate the consequential aspect of this research programme. This topic is addressed in the next section of this chapter.

6.4 Monte Carlo Simulation

As elaborated on in Chapter 2, one of the six aspects of the Messick (1995) approach to the construct validity of personality measures in selection situation is that of the consequential effects - "The consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice" (p. 745). The theoretical background and empirical evidence of the consequential aspect (Messick, 1995) of socially desirable responding in the form of faking good in high stakes employee

selection contexts were reviewed in Section 5.2 of the previous chapter. This section of Chapter 6 deals with the methodological solution used in this research programme to take account of the consequential aspect of Messick's (1995) concept of construct validity. Bäckström, Björklund, and Larsson (2009) point out that the pressure of socially desirable responding is likely to be stronger in an applied context, such as that of the field studies of this research programme, compared with basic research studies. As a result, the full effect of CMV due to social desirability is often underestimated when calculating relationships affecting criterion related validity, according to the authors. They also emphasise that social desirability is likely to be caused by situational pressure such as in found in recruitment studies, the subject of the research programme.

This aspect of CMV due to socially desirable responding has been examined and explored in detail earlier in Chapter 4 but is worth repeating and emphasising again. Therefore it is very important that the consequential aspect of the use of a personality measure, such as the NEO PI-R, be investigated as comprehensively as possible in order to carry out a proper construct validity (Cronbach & Meehl, 1955; Embretson, 1983; Embretson, 2007; Loevinger, 1957; Messick, 1995; Smith, 2007; Strauss & Smith, 2009) evaluation of the use of the NEO PI-R in high stakes employee selection contexts. It was not possible to directly investigate this aspect of construct validity directly in the field studies.

Such research would have been difficult to conduct in an ecologically valid manner even using traditional experimental or quasi-experimental techniques, and so in order to overcome this problem in fully evaluating the construct validity of the NEO PI-R a Monte Carlo simulation approach was used. Monte Carlo simulation is a type of simulation that relies on repeated random sampling and statistical analysis to

compute the results. This method of simulation is very closely related to random experiments, experiments for which the specific result is not known in advance. In this context, Monte Carlo simulation can be considered as a methodical way of doing so called 'what-if' analysis (Raychaudhuri, 2008). It has been used in a similar manner by Murphy and Shiarella (1997) in estimating the validity of general cognitive ability tests and personality tests in predicting 'job performance', where performance is conceptualized as a composite of multiple performance measures, and by Komar, Brown, Komar, and Robie (2008) to investigate the criterion related validity effect of faking good on the prediction of job performance. The technique was also used by other researchers for investigating the effect of socially desirable responding on the criterion related validity of personality traits (Berry & Sackett, 2009; Converse, Peterson, & Griffith, 2009; Paunonen & LeBel, 2012; Marcus, 2006).

Every Monte Carlo simulation starts off with developing a deterministic model which closely resembles the real scenario, and which takes the random values of the input variables, and transforms them into the desired output using the model. The model is then simulated repeatedly by repeated sampling from the pre-specified probability distribution of the input variables, so that a probability distribution of the outputs of interest is obtained. This part is the core of Monte Carlo simulation (Raychaudhuri, 2008). The value of each output parameter is one particular outcome scenario in the simulation run. The output values from a number of simulation runs are collected and aggregated, by means of the appropriate statistical analysis on the values of the output parameters. This step provides researchers with statistical confidence for the conclusions or inferences arrived at after running the simulation. The sampling statistics of the output parameters are used to characterise the output variation. Obviously as the number of simulation trials increases the accuracy of the

estimates of the expected value of the outputs variable increases (Paxton, Curran, Bollen, Kirby, & Chen, 2001; Spence, 1983; Raychaudhuri, 2008).

The outputs from the repeated simulations are treated in exactly the same way as repeated experiments in any setting (Spence, 1983). Averaging trial output values result in an expected value of each of the output variables. Aggregating the output values into groups by size and displaying the values as a frequency histogram provides the approximate shape of the probability density function of an output variable. The output values can themselves be used as an empirical distribution, thereby calculating the percentiles and other statistics. Alternatively, the output values can be fitted to a probability distribution, and the theoretical statistics of the distribution can be calculated. These statistics can then be used for developing confidence intervals around the output variables (Raychaudhuri, 2008). By using a Monte Carlo simulation approach it will be possible, in this research, to investigate the likely incidence of selecting a candidate who has faked good on the NEO PI-R, the personality measure used in the field study. Without this approach, it would not have been possible to fully evaluate the construct validity (Messick, 1995, Embretson, 2007) of the NEO PI-R in a high stakes employee selection situation.

It should be emphasised that the outcomes of any Monte Carlo study depend heavily on the range of parameter – inputs and outputs - values studied (Paxton et al., 2001; Spence, 1983). Both cognitive ability and the Big Five dimensions, on both a bivariate and a multivariate basis, have been linked to the construct of overall job performance. For example, the link between cognitive ability and individual job performance is one of the most widely studied topics in psychology. Murphy and Shiarella (1997) showed that in studying job performance the mean validities obtained using different combinations of predictor and criterion related constructs can vary

extensively. In their Monte Carlo investigation, to fully evaluate the link to the job performance domain they used different combinations of measures of the constructs of task performance and organisational citizenship behaviour.

It is therefore important that in Monte Carlo simulations of the outcomes of high stake selection, using personality measures, that different combinations of predictors of job performance be tested because of what Murphy and Shiarella (1997) found - "The construct of job performance is one that is defined by the demands of the job, the structure, strategy and mission of the organization, and so forth, and jobs that are similar in terms of their titles, main duties, and so forth may still yield very different definitions of what constitutes good or poor performance" (p. 844).

In using the Monte Carlo simulations to examine the consequential effects the mathematical property that Dawes (1979) referred to as the flat maximum effect, whereby linear model regression weights that are near to optimal lead to almost the same output as do optimal regression beta weights, was taken into account. Wainer (1976) subsequently showed mathematically that what he termed 'the equal weights theorem' (p. 214) proves that an actuarial prediction using linear weights is apt to be very close to the optimal one, were the optimal weights known, and often better than one which does not use optimal weights, provided that (a) all predictor variables are oriented in the proper fashion, discarding equivocal ones; and (b) scaling them all into standardised form. Wainer (1976) also showed that this approach works well even when an operational criterion is not available, as is often the case in senior executive selection situations. Independently of Wainer (1976), Einhorn and Hogarth (1975) also showed mathematically that unit weighing schemes for predictors will usually be a more than satisfactory approximation to all the possible optimal weighting schemes, even where the criterion could not be defined. Dawes (1979) as well as showing that

improper (i.e. equal weights) linear models performed well when predicting criterion outcomes, addressed the main technical, psychological and ethical objections to the use of improper linear models. These mathematical properties of linear regression were used in order to test a number of different scenarios in the Monte Carlo simulations of this research programme.

The preceding sections of this chapter have set out in detail the basis for the methodological approaches used in addressing the core objectives of the research programme, namely, establishing the construct validity of the NEO PI-R in high stakes selection situations. As McKenzie et al. (2011) point out, construct validation procedures often “underutilize techniques that provide evidence that the set of items used to represent the focal construct actually measures what it purports to measure” (p.293). In an earlier article MacKenzie (2003) stated that “the problems of poor construct validity and statistical conclusion validity that plague many manuscripts can be minimized if you carefully define the focal constructs in your research, make sure that your measures fully represent them, correctly specify the relations between the measures and constructs in your measurement model, and stick to it” (p.326). The advice of MacKenzie was followed in this chapter.

To summarise the chapter contains a review of a number of different methodological issues were examined in detail. This showed that extant MTMM studies provided an empirical basis for the conclusion that there is evidence for two higher order factors, but no valid evidence for a GFP. It was also shown that the results of factor analysis, particularly CFA, must be fully understood and used with caution before arriving at adequate conclusions. In addition, it was noted that CFA has a research useful role to play because it can be used for mathematical modelling. Relying primarily on the recent Connelly and Chang (2016) meta-analytic research it

was argued that a case can be made for the adequacy and appropriateness of the bespoke BIDR-IM measure use in the research programme to detect faking good. The use of the Monte Carlo simulation technique allows the consequential aspect of construct validity to be investigated. As a result of these factors, it is now possible to state the research hypotheses tested in the research programme.

6.5 The Research Hypotheses Tested

The methodological considerations concerning the MTMM approach to separating trait from method effects (Chang, Connelly, & Geeza, 2012), the proper use and a full understanding of factor analysis in model testing (McDonald, 1999), the construct validity of impression management measures, and the assessment of job performance effects in the absence of empirical data, all impact on the achievement of the objectives set out in Chapter 1 of this thesis. Without a thorough understanding of these four issues, it would not be possible to fully answer the questions originally posed in Chapter 1, and reviewed in the succeeding chapters. The preceding chapters to this one have set out, in detail, the background understanding which was necessary for what now follows in the remaining chapters of this thesis. Following Embretson's (2007) division of the process of establishing construct validity the methodological key to determining the 'internal' meaning aspect of construct validity of the NEO PI-R in the research programme was CFA. The use of the bespoke BIDR-IM measure as a method for dichotomising faking was of major importance in determining the 'external' aspects of its construct validity, which is why the topic was explained and examined in such detail in this chapter.

The methodology reviewed in the previous sections established the parameters for defining the hypotheses to be tested in the research programme. Based on the research evidence reviewed in Chapters 2 to 5 of this thesis in order to establish the internal aspect of construct validity of the NEO PI-R in high stakes employee selection situations the following competing hypotheses, which putatively could arise due to a higher order factor structure superordinate to the Big Five, were proposed and tested:

(1) there are two uncorrelated higher order level factors, superordinate to the Big Five, which are not methodological artefacts and which are based on Digman's (1997) research (H1);

(2) there is a single higher-order factor, the General Factor of Personality (GFP) superordinate to the Big Five (H2) (Rushton & Irwing, 2008) (H2).

A hierarchical model of personality with the two latent factors Plasticity and Stability loading on a GFP was not tested because as Finch and West (1997) point out and, as mentioned before but worth repeating, Kline (2011) makes clear when he states that "there must be at least three first-order factors. Otherwise, the direct effects of the second-order factor on the first-order factors or the disturbance variances may be underidentified" (p. 249). This is also the view of McDonald (1999, p.188). It is important to note here that if Hypothesis 1 (H1) is correct then a standard congeneric CFA model of the Big Five with two indicators loading on Plasticity should not yield a solution because of underidentification. This has important implications for

determining whether the formal warning was successful in, at least, minimising faking good in the field studies of the research programme.

There is also another hypothesis concerning the higher order structure that ideally should be tested (Anusic et al., 2009; Ashton, Lee, Goldberg, & de Vries, 2009; Biderman, Nguyen, Cunningham, & Ghorbani, 2011) which is that any higher order factors detected were solely due to these other methodological and statistical artefacts, rather than shared variance of the Big Five, could not be directly tested in this research programme because of the existence of the previously established secondary factor loadings of items in the personality measure used (Ashton et al., 2009; Hofstee, de Raad, & Goldberg, 1992; Hopwood, Wright, & Donnellan, 2011; Johnson, 1994; Saucier 2002). In addition, it was not possible to fully control for the effect of other artefacts such as the evaluative content of item wording in the personality measure used (Bäckström et al., 2009; Biderman et al., 2011), or acquiescence bias among participants (Anusic et al., 2009). This has implications arising from correlated method factors (ρ_{M1M2}) in equation E2 on page 136 of this chapter in that a correlation between the two putative higher order factors could exist solely because of these other method factors that were not controlled for in the research programme.

The methodological key to establishing the construct validity of the use of NEO PI-R omnibus personality inventory for high stakes employee selection purposes – the primary objective of this research - depends on the answers to the research questions posed by H1 and H2. If H1 turns out to be correct and H2 false then it can be inferred that the use of the formal verbal warning about the assessment containing measures to detect impression management did eliminate or minimise lying by participants in the form of faking good. Such a finding would be consistent with the

extant MTMM studies reviewed in subsection 6.1.1.1. In addition, the use of the bespoke BIDR-IM scale will help to identify the extent to which any faking good that occurs, in spite of the formal warning, can lead to biased or unfair selection decisions. Both of these issues are extremely important in the applied employee selection setting.

The remaining chapters apply both the theoretical concepts explored in detail in Chapters 2 to 5, as well as the methodological guidance of this chapter, to the results and analyses of the field studies of the research programme. The next chapter details how the methods were actually used, together with procedures followed in the research programme.

Chapter 7

Research Methods

This chapter contains details, in Sections 7.1 and 7.2, of the two field study samples of participants and the measures used in the research programme. Section 7.3 describes the test administration procedures followed. In Section 7.4 the procedural steps followed in using the various analytical techniques employed are described. This research programme made use of a range of analytical tools. To assess the structural aspect of construct validity an Exploratory Factor Analysis and a Confirmatory Factor Analysis were conducted. To investigate the generalisability aspect Multigroup Invariance Analysis was employed. To explore the consequential aspect and the dichotomisation of participants in the Managerial field study Cluster Analysis and Monte Carlo Simulation were used. These different analyses were carried out in order to establish a strong (Kane, 2001) construct validity case for the use of the NEO PI-R in high stakes employee selection situations.

The theoretical and empirical evidence rationales for this approach were provided in Chapter 6, and the approach followed in the analyses carried out is also consistent with the aspects of strong construct validity as enumerated by Messick (1995). Recent reviews of the construct validation process (Kane, 2013; McKenzie, Podsakoff, & Podsakoff, 2011; Strauss & Smith, 2008) support the principles of construct validation. The reliance on multiple forms of empirical evaluation of the nomological net of the psychometric measures used in this research programme is consistent with Messick's (1995) definition of the construct validation process which explicitly calls for the use of multiple forms of evidence.

7.1 Participants

The primary field study contained 443 participants ('Managers') all of whom were applicants for a wide range of middle and senior management positions in a range of companies. They were all individually assessed for different client companies. 29.8% of the participants were female. Their average age was 38.5 ranging from 26 to 59, with a standard deviation of 8.5. The positions for which the participants were being considered included CEO, CFO, COO, and various other middle and senior technical, financial, operational and sales/marketing management positions in a number of diverse organisation. The organisations involved covered a wide range of sectors such as energy, transportation, distribution and wholesaling, manufacturing, NGO's, consulting, and construction.

The validation field study consisted of 201 applicants for senior positions in a large company, all of whom completed a version of the NEO as part of the selection process. There was very limited demographic data available from the commercial organisation that provided the data for the applicants, other than their names.

7.2 Measures

All the participants in the Managerial sample completed the same battery of tests as part of an individual assessment, which included two ability measures, an omnibus personality measure, and an impression management measure.

Personality. All participants in the Managers field study completed the paper and pencil version of the NEO PI-R (Costa and McCrae, 1992) personality measure. The test administration was proctored. This measure is a widely used, commercially available, omnibus measure which contains 240 items measuring the five broad Big Five personality dimensions. The measure contained 48 items for assessing each of the Big Five. The NEO PI-R is available in a number of languages and has been the subject of an extensive body of research in different settings and different countries (McCrae and Antonio, 2005). The broad dimensions are further divided into 6 facets each being measured by 8 items. The items were scored using a five point Likert scale with five response options each scored either 0,1,2,3 or 4 depending on the response. Participants were instructed to read each item carefully and to circle the one answer that best corresponded to their agreement or disagreement with the item in question. The response options were ‘Strongly Disagree’ (SD), ‘Disagree’ (D), ‘Neutral’ (N), ‘Agree’ (A), or ‘Strongly Agree’ (SA). The participant was instructed to respond SD if the statement was ‘definitely false’ or if they strongly disagreed; D if the statement was ‘mostly false’ or if they agreed; N if the statement was about equally true of false, if they couldn’t decide, or if they were neutral; A if the statement was ‘mostly true’ or if they agreed; and SA if the statement was ‘definitely true’ or if they strongly agreed.

The cover page for the NEO PI-R questionnaire was altered to remove all reference to the NEO PI-R being a ‘Personality Inventory’, because of socially desirable responding concerns arising from the research reviewed in Chapters 3 and 4. Every second item in the questionnaire is reversed scored. Positively-keyed items are items that are phrased so that an agreement with the item represents a relatively high level of the attribute being measured. Negatively-keyed items are items that are phrased so that an agreement with the item represents a relatively low level of the

attribute being measured. The reason for alternating item wording is to minimise extreme response bias and acquiescent bias (Paulhus, 1984). According to Murphy and Davidshoffer (1998), if equal numbers of positive and negative responses are keyed on a test, then any tendency to acquiesce or be critical will not influence the test score markedly. The scale scores of each of the Big Five dimensions were obtained by summing the item score totals for each facet. These scores were used as the input for the factor analyses and Monte Carlo simulations.

The participants in the Validation field sample all completed the on-line version of the NEO-PI3. The NEO PI-3 was used instead of the NEO PI-R because it was the version that was available which allowed for remote on-line assessment of job applicants. The NEO-PI3 is a revision of the NEO PI-R. The NEO-PI3 retains the reliability and validity of the NEO PI-R and, when introduced, featured new normative data (McCrae, Costa, & Martin, 2005). According to McCrae et al. (2005), the NEO-PI3 eliminated most of the items in the NEO PI-R that adolescents aged 14 to 20 find difficult. Also according to McCrae et al. (2005) the NEO-PI3 shows modest psychometric improvements over the generally good performance of the NEO-PI-R. The measure was administered online and was unproctored.

Cognitive Ability. Two cognitive ability measures were used in the Managers field study – the AH4 (Heim, 1970) and Raven's Advanced Progressive Matrices (Raven, 1965). The AH4 test is a widely used standard test of intelligence or cognitive ability measure that consists of two parts – Part One assesses the verbal and numerical content domain, and the other part assesses the figural content domain (Duncan, Seitz, Kolodny, Bor, Herzog, Ahmed, Newell, & Emslie, 2000; Heim, 1970). Raven's APM is a non-verbal test designed to be a culture free measure of abstract reasoning ability that does not rely on crystallized knowledge (Carpenter, Just, & Shell, 1990;

Prabhakaran, Smith, Desmond, Glover, & Gabrieli, 1997). Both the AH4 and Raven's APM have been shown to have high reliability. Both tests were timed tests – each part of the AH4 measure lasted 10 minutes, and the APM lasted 40 minutes – and the tests were scored on the basis of the number of items correctly answered. The Validation sample did not complete any ability measured when completing the NEO-PI3.

Impression Management. As described earlier in Chapter 5, Paulhus (1998) developed a measure of socially desirable responding, the Balanced Inventory of Desirable Responding (BIDR), which assessed two dimensions – Impression Management (IM) and Social Deception Enhancement (SDE) - (Paulhus, 1984; Paulhus, 1998; Paulhus & Reid, 1991; Paulhus, Harms, Bruce, & Lysy, 2003). IM refers to an intentional distortion of self-descriptions in order to be viewed favourably by others. SDE, in contrast, denotes an unconscious propensity to think of oneself in a favourable light. As explained in Chapter 5 Paulhus's (1984) distinction between the two dimensions of socially desirable responding can help to shed light on whether the putative higher order factors of personality exist or not, in the case of deliberate faking good better than other measures of socially desirable responding which do not distinguish between these two dimensions, because of the volitional nature of IM compared with SDE (Barger, 2002; Lonnqvist, Irlenbusch, & Walkowitz, 2014).

The IM scale contains 20 items, half of which are reversed scored. The items in the IM scale are both overt and clear cut. As explained in Section 6.3 of the previous chapter, in order to deal with a concern for item transparency rather than grouping them together the 20 IM items were randomly embedded in a bespoke questionnaire of 75 items containing the 20 IM items and 55 distractor items. The distractor items were taken from Button's Goal Orientation measure (Button,

Mathieu, & Zajac, 1996) and Judge's Core Self Evaluations measure (Judge, Erez, Bono, & Thoresen, 2003).

Scoring of the bespoke BIDR-IM was based on a seven point Likert scale with the extreme response option corresponding to 'not true' which was scored 1 on the Likert scale, and the extreme response option 'very true' scored as 7. Response options for responses between the extremes were scored from 2,3,4,5 and 6 corresponded to a greater or lesser degree of 'somewhat true'. Participant responses on the completed questionnaire at 6 or 7 on the 7 point Likert scale were then given a score of '1' on positively worded items, whereas responses at 1 or 2 on the 7 point scale were scored as '1' on negatively worded items. All other responses were scored as '0'. The maximum score that could be obtained is 20 and the minimum score is zero.

7.3 Procedure

The two measures of primary interest – the NEO PI-R and the bespoke Impression Management measure - were part of the battery of tests administered during the same session to each of the Managerial sample participants, which also included the two ability tests, as mentioned earlier. The tests were administered in the course of a single assessment session. The order of test administration was the AH4 cognitive ability test, followed by the NEO PI-R, the Raven's ability test and, finally, the bespoke IM measure. The bespoke IM measure and the NEO-PI3 were administered online for the Validation sample.

The battery of tests was administered on an individual participant basis for the Managerial sample and the testing was proctored by the same person for all

participants. A formal verbal warning in the form of a statement “*The test battery that you are completing contains measures to detect if there is any element of deliberate impression management in your responses. This is to ensure that an accurate assessment of you is obtained*” was given to each individual assessed in the Managerial sample used in this study at the end of the instructions given at the start of the personality assessment.

Each participant in the Validation field study was required to read what was entitled a psychometric honesty statement on a computer screen and to tick a box acknowledging that they had read and understood the statement. The statement was presented on screen to the applicants before they completed the NEO-PI3 and then the bespoke BIDR-IM scale on line. The statement contained the following warning:

“There is a natural tendency among job candidates to mistakenly try to create a favourable impression. This can invalidate the testing because all job candidates are human and as such, not perfect. In order to minimise the possibility that applicants who provide inaccurate or dishonest responses when answering the items are hired, we use a special scoring system. More specifically, the tests used are designed to detect if you attempt to lie in your responses. They detect dishonest answers and this will negatively affect your scores and will not increase your chances of getting the job.”

7.4 Analyses

The results of primary interest from the test administrations were the summed raw scale scores for each of the Big Five dimensions, which were subjected to a

number of different analytical procedures – Exploratory and Confirmatory Factor Analysis, Cluster Analysis, Multigroup Invariance Analysis, and Monte Carlo simulations – which are now described in turn.

7.4.1 Factor Analysis

The participants' raw scores were initially subjected to an Exploratory Factor Analysis (EFA) based on principal components analysis and oblique rotation, using SPSS Version 21. Principal component analysis was used because the factor loadings in Table 5 of the NEO PIR Professional Manual were obtained using principal component analysis (Costa & McCrae, 1992; Hopwood & Donnellan, 2010), and because of the likely presence of specific variance due to correlated errors (Costa & McCrae, 1995; Johnson & Ostendorf, 1993;). It was expected, based on previous research detailed earlier in Chapter 6, that a higher order factor structure consisting of two latent factors would emerge. Conscientiousness, Agreeableness, and Emotional Stability (Neuroticism reversed scored) were expected to load on one factor, and Extraversion and Openness were expected to load on the second factor (Aluja, García, García, & Seisdedos, 2005).

Following this, the results were subjected to a Confirmatory Factor Analysis (CFA) using AMOS Version 23. Analyses were carried out using the maximum likelihood estimation method for estimating the parameters of the models tested. An Unmeasured Latent Method Factor (Biderman, Nguyen, Cunningham, & Ghorbani, 2011; Johnson, Rosen, & Djurdjevic, 2011; Podsakoff, MacKenzie, & Podsakoff, 2012; Podsakoff, MacKenzie, Lee, & Podsakoff, 2003) was included in a number of the models tested in the CFA's carried out in order to detect the presence, or

otherwise, of a common method factor. This would enable common method variance (CMV) to be modelled by including factor loadings from the putative latent CMV method factor to all of the indicators of both Plasticity and Stability in the relevant models tested (Podsakoff et al., 2003). This technique does not require the researcher to measure the specific factor responsible for the method effect (Johnson et al., 2011; Podsakoff et al. 2003). This approach was taken, also, because a number of monomethod studies had previously found evidence for the existence of a sixth factor, in addition to the Big Five, which has been attributed to method effects (Cellar, Miller, Doverspike, & Klawnsky, 1996; Khele et al., 2012; Lim & Ployhart, 2006; McFarland & Ryan, 1993).

A priori it was expected that there would be some correlated residuals because, as already described in detail in Chapters 3 and 6, many of the items in the NEO PI-R are likely to have a secondary loading on a second Big Five factor as well as a primary loading on a different one of the five factors (Costa & McCrae, 1995; Hofstee, de Rand, & Goldberg, 1992; Hopwood & Donnellan, 2010; Johnson & Ostendorf, 1993; Johnson, 1994). For this reason the modifications indices (Brown, 2006; Kline, 2011) of the CFA models tested were examined as part of the model respecifications that were carried out. Some of the CFA models were re-tested having made theoretical justifiable adjustments to the model for the presence of correlated errors, as indicated by the modification indices of the AMOS analysis (Byrne, 2010).

7.4.2 IM Cut Off Score

Following the recommendation of Paulhus (1998), and the construct validity analysis of the BIDR-IM scale of Section 6.3 of the previous chapter, participants

scoring 12 or higher on the IM scale was deemed to be purposely self enhancing. The distribution of IM scores was dichotomised for most of the analyses carried out based on a faking good cut-off score of 12 because it was assumed that faking good had occurred with scores of 12 or higher. The manual for the BIDR (Paulhus, 1998) regards scores of 12 or higher as ‘probably invalid’ (p. 12). Ellingson, Heggstad, and Makarius (2012) used a BIDR IM scale score of 6 as the predetermined criterion for flagging participants that should be retested because of a faking good concern. This criterion was selected by Ellingson et al. (2012) for two reasons. Firstly, it was above the mean of score distributions on the BIDR–IM observed in samples of participants responding under neutral conditions and, secondly, it was the median value observed on the BIDR–IM in pilot testing of the Ellingson et al. (2012) study manipulation. Fan, Gao, Carroll, Lopez, Tian, and Meng (2012) found in their study that the mean BIDR-IM scale score after a warning was 11.93 on the BIDR-IM used in their research, for the participants flagged as possible fakers. Because of this latter study in which possible fakers who were flagged were retested, as well as the Paulhus ‘probably invalid’ score of 12 it was decided to use 12 as the cut-off score for dichotomising the BIDR IM scores into fakers and non-fakers. The research, described earlier in Chapter 4, of Fischbacher and Heusi (2008) together with the mathematical formulation of Borghans, Duckworth, Heckman, and ter Weel (2008), described in some detail in Chapter 6, also supports the decision to use a cut-off score of some particular value as the basis for dichotomisation. The choice of a cut-off score of 12 represents an arguably acceptable balance between a desire to obtain a dichotomisation that was defensible from a construct validity perspective and the absence of certainty. To investigate this matter further a Cluster Analysis was carried out to see if there was empirical support in the Managerial field study sample for the

use of a cut-off score of 12 to dichotomise the participants in the Managerial field study sample.

7.4.3 Cluster Analysis

There is evidence from other research that participants in research studies of the incidence of faking good in personality assessments cluster into a small number of groups (Hauenstein, Bradley, O'Shea, Shah, & Magill, 2017; Robie, Brown, & Beaty, 2007; Zickar, Gibby, & Robie, 2004). The analytical technique of Cluster Analysis (Milligan & Cooper, 1987) was used to see if meaningful clusters emerged in this research programme from the Managerial field study. The Managerial field study data were examined on a post hoc exploratory basis using Cluster Analysis to see if the participants could be meaningfully grouped based on the recent findings of Connelly and Chang (2016) with respect to the factors underlying socially desirable responding on self-report personality measures. As described in Chapter 6 they found in their MTMM meta-analytic investigation of social desirability scales, including the BIDR IM scale, which the IM scale is accounted for by both self-report method variance (style) and trait factors (substance). As well as being meaningfully associated with self-report method variance the IM scale, in the Connelly and Chang (2012) research, also had an association with Conscientiousness and Agreeableness.

There is no distinction made in cluster analysis between independent and dependent variables. Therefore the exploratory clustering exercise carried out was based on participants' scores on three of the measures used in assessing participants' personality – the bespoke BIDR-IM measure, Agreeableness, and Conscientiousness. The clustering procedure followed was that used by Thiele, Kubacki, Tkaczynski, and

Parkinson (2015). SPSS 21 was used to perform a two-step cluster analysis using the log-likelihood procedure (Bacher, Wenzig, & Vogler, 2004) to see if groupings based on IM, Agreeableness and Conscientiousness scores would form meaningful clusters in the data set of the field study. These putative groupings - if any meaningful ones are found - could help in the construct validity investigation of the bespoke BIDR IM scale as well as the choice of the dichotomisation cut-off score of 12, which was used in this research programme.

Following Santos and Horta (2015), the procedure involves two steps. In a pre-clustering step, a modified cluster feature (CF) tree is constructed by means of sequentially scanning each case on the database and deciding whether it should be incorporated into a pre-existing “branch” of the tree or assigned to a new one. The construction of the CF tree is based on an algorithm that selects a number of pre-clusters that are used in the subsequent step. An outlier selection procedure is then carried out. The second step is the clustering itself. This step, involves running a hierarchical agglomerative analysis (Santos and Herzog, 2015). The pre-clusters are used instead of the individual cases as a way of overcoming issues that typically arise from using a hierarchical cluster analysis, such as computational limitations or extreme partitioning of the data. This step uses an auto-clustering algorithm, which can be used to determine the optimal number of clusters without having to resort to subjective measures, such as graphic interpretations, which are often used in traditional methods. Model fit is evaluated through the average silhouette measure of cohesion and separation. This represents the average of all cluster silhouette measures, ranging from -1 to 1 and is used to measure the relative cohesion (positive values) of data points in a given cluster or, inversely, their separation (negative values).

7.4.4 CFA Invariance Analysis

The generalisability aspect of Messick's (1995) approach to establishing construct validity involves an examination of the extent to which test score properties and interpretations apply to different populations and groups. An invariance validation exercise of the findings of the field study of job candidates in the Managerial sample was carried out by administering the NEO-PI3, together with a formal warning, to a different group of job candidates (the Validation sample). The CFA results of models evaluated in both field study samples were then tested for invariance in the two groups of job candidates, following the generally accepted procedure for determining the factorial equivalence of a measure in different groups (Bagozzi & Edwards, 1998; Brown, 2006; Byrne, 2010; Ion & Iliescu, 2017; Kline, 2011; MacKenzie, Podsakoff, & Podsakoff, 2011; Vandenberg & Lance, 2000). The results of the CFA models were tested, using the multiple group comparison procedure of AMOS, by comparing the CFA model parameters in both the Managerial and Validation field study samples. The procedure allows for comparisons to be made between the two field study samples parameter estimates and by using a Chi Squared difference test to determine whether or not there is a statistically significant difference between the target and validation groups (Vandenberg & Lance, 2010). Invariance is deemed to have been detected when no significant differences are found between the groups.

Following the generally accepted procedures, detailed in Byrne (2010) for testing for invariance using AMOS 23, construct validity was tested by, firstly, testing for configural (similar factor and indicator configuration) invariance. The unconstrained configural model incorporates a comparison of the sample data of the

combined Managers' and Validation' samples, assessed simultaneously by AMOS 23, with the hypothesised CFA model. This multigroup model then allows for the parameters of both samples to be estimated at the same time (Brown, 2010). This was followed by testing for measurement invariance with respect to metric invariance (equality of factor loadings), then testing for structural invariance (equality of factor variances and covariances), followed by testing residuals for invariance. Full measurement invariance, i.e. strong construct validity, is established when equality of error variances, factor loadings, and factor covariances, if any, is shown to exist between the groups tested (Chan & Schmitt, 1997). Homogeneity of the two groups was also tested by including a scalar test of measurement intercepts in the analysis (Brown, 2006).

7.4.5 Monte Carlo Simulations

As described in Section 6.4 of the previous chapter using different unit weighted combinations of a number of different job performance predictors - cognitive ability and personality traits - the question of how often are applicants who have faked good can be selected for management positions under different selection criteria was investigated. This approach has already been used by others (Converse, Peterson, & Griffith, 2009; Komar, Brown, & Komar, 2008, Paunonen & LeBel, 2012) to explore the effect of faking good on selection by using, for example, cognitive ability predictors to partly offset the effect of faking good in personality measures. A series of Monte Carlo simulations was performed in order to examine the proportion of small subsets of finalists, drawn from the real world executive selection applicant pool of the Managers field study participants, which contain candidates who

were deemed to have faked good because they scored high on the bespoke BIDR-IM impression management measure.

Using the random number generator function in Microsoft Excel two thousand simulations were run to select a short list of candidate finalists, for a scenario based on three, four or five finalists, from the population of the participants. Each of the randomly selected set of finalists was examined to see if any one of the selected sets of three (or four, or five) in the simulation contained a finalist with an IM score of 12 or higher. Two statistics which were produced by the simulations were examined – the proportion of sets of finalists with a least one ‘faker’ (defined as having an IM score of 12 or higher), and the proportion of selection decisions from the finalists sets in which the ‘faker’ was selected for the hypothetical position. The overall proportion of those sets with a ‘faker’ was calculated as a percentage of the 2000 simulated sets. The simulation outcomes were based on the modelling pre-condition that only one finalist out of each set would be selected (or win).

Following this a number of compensatory weighted linear models, using standardised scores, were used as the basis for selecting the ‘winning’ finalist from the sets of 3, 4 or 5 finalists. Cognitive ability measures are faking good resistant so the baseline model was selection based on cognitive ability alone (AH4). The first comparative model tested was based on participants’ Conscientiousness scores, which is what Peterson, Griffith and Converse (2009) used in their study together with a cognitive ability measure. The second model tested used a unit weighted composite of Conscientiousness, Neuroticism and Extraversion. The third model tested used a unit weighted model of cognitive ability, Conscientiousness and Neuroticism. The fourth model consisted of a linear model with equal weighting for Conscientiousness and Neuroticism. The fifth model with the weighting for cognitive ability 2.5 times that of

Conscientiousness and Neuroticism. The final model tested was the same as Model 5 with the addition of Extraversion and Openness.

The rationale for using a model with a greater weighting for cognitive ability was based on Einhorn and Hogarth's (1975) point that "an investigator may, of course be able to inject more specific prior information into the analysis" (p. 189). The relative importance of psychometric 'g' in determining job performance increases with increase job complexity (Gottfredson, 1997; Schmidt & Hunter, 2004). A relative weighting of 2.5 used for cognitive ability was based on Schmidt and Hunter's (2004) meta-analytic finding. The inclusion of personality dimensions in the different models was based on the findings of a wide range of studies of the criterion validity of the Big Five (Judge, Higgins, Thoresen, & Barrick, 1999; King, Walker, & Broyles, 1996; Ones, Viswesvaran, & Dilchert, 2005; Salgado, Moscoso, & Lado, 2003).

7.4.6 Comparison with the Rosse, Stecher, Miller, and Levin (1998) Study

The final analysis carried out was a comparison with the one extant study using both the NEO PI-R and the BIDR-IM in an applied study of both job applicants and job incumbents. To evaluate the extent of faking good it was not possible in this research programme to perform either a between participants or within participants evaluation of the two conditions – job applicant and/or job incumbents - as the research programme of this thesis only dealt with convenience field studies samples of job applicants. In order to partly overcome this problem the results of Rosse, Stecher, Miller, and Levin, (1998) were used as the source for a job incumbent sample

for comparison purposes, which was not available in either of the field studies of this research programme. The procedure followed in their research used both the NEO PI-R and the IM scale from the BIDR. In using the BIDR-IM Scale they used the approach of random insertion of the 20 items in the personality measure. As a form of further construct validity check (Messick, 1995) of the generalisation aspect, and the impact (Embretson, 2007), the average scores on the Big Five Dimensions for the participants in the Managerial field study were compared with the average scores for Rosse et al.'s job applicants and incumbents. If the Rosse et al. (1998) job incumbents' scores and those of the Managerial field study were found to be of the same order of magnitude this then could be taken as further support for the effectiveness of the formal warning in this research programme. In addition, the average scores for norm group of the BIDR IM scale in the publisher's manual were compared with the IM scores for participants in the Managerial field sample. Cohen's 'd' effect sizes and the corresponding 95% confidence intervals were of this comparison were used to also help in the evaluation of the research question as to whether faking good had been, at least, minimised or not.

To summarise the contents of this chapter contain a review of the participants, methods, procedures and analytical techniques used in the research programme. The chapter was structured so as to outline the overall procedural approach followed for the utilisation of analytical techniques described in Chapter 6 as well as the use of the techniques of Cluster Analysis and Multigroup CFA Invariance Analysis in the analysis of the data. This approach is an essential part of Messick's (1995) 'overall evaluative judgment' (p. 741) approach to construct validity that would bring together the theoretical, conceptual and analytical issues, relevant to the research programme and detailed in Chapters 1 to 6, and the data obtained in the two field studies. The

next chapter provides details of the results of the different analyses carried out using the data from the two field studies.

Chapter 8

Results

This chapter reports and summarises the results of the different data analyses carried out as part of the research programme. The analyses followed the approach to establishing the evidential and inferential bases for strong construct validity as set out in a number of articles (Embretson, 2007; Messick, 1995; Smith, 2005) already referred to in Chapter 2. These analyses build on the construct validation objective of meeting the requirements of Messick's (1995) six aspects of construct validation as explained in Chapter 6.

The flow chart of Figure 7 below illustrates the steps carried out to establish strong construct validity in this research process which followed Messick's (1995) six aspects. Specifically, Section 8.1 of this chapter reports the descriptive statistics and analyses for the Managerial field study sample. Sections 8.1.2 and 8.1.3 provide the results of the exploratory and confirmatory factor analyses of the Managerial field study sample are provided. This covers both the substantive and structural aspects of Messick's approach. The results of a CFA invariance validation comparison between the Managerial and the Validation field study samples are also provided in Section 8.1.3, in order to meet the generalisation aspect of Messick evidential basis. These sections are central to proving or disproving the hypotheses of Chapter 7. Next I the results of the Cluster Analysis are reported in Section 8.2 which was carried out as part of the construct validation process for the bespoke BIDR-IM measure, and which was necessary for providing some of the analytical evidence in this research

programme that was necessary for establishing the construct validity of the measure.

The next Section, 8.3, contains the results of the Monte Carlo simulations which were carried out in order to comply to some extent, in the absence of job performance data, with the consequential aspect of Messick’s (1995) construct validation requirements.

Finally, the results of a comparison with an extant study using the NEO PI-R and the bespoke BIDR-IM are shown in Section 8.3.3.

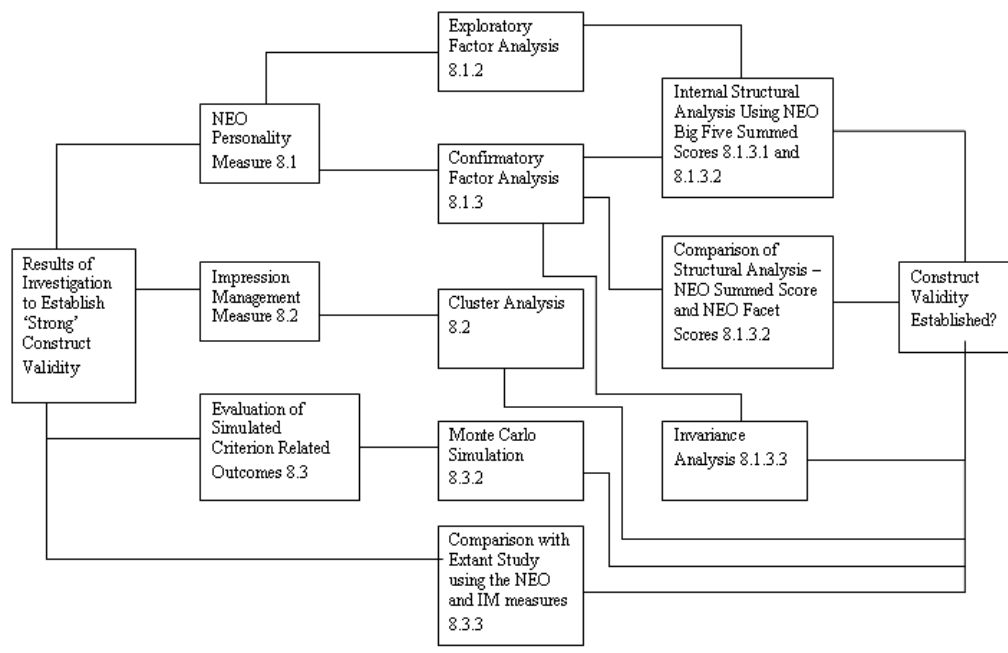


Figure 7 Flow Chart for Results Sections of Chapter 8

8.1 Managerial Field Study

8.1.1 Descriptive Statistics

This section contains the descriptive statistics for the Managerial sample. It also includes a comparison of the descriptive statistics for the two dichotomised bespoke BIDR-IM groups, and the correlations between the Big Five dimensions in the sample. Table 3 contains details of the descriptive statistics for the participants in the full Managerial field study sample of 443 participants. With the exception of ‘Agreeableness’ the mean scores for the full sample are all substantially higher (lower in the case of Neuroticism) than those for the norming groups in the test publisher’s manual of the NEO PI-R (Costa & McCrae, 1992). The effect size, as measured by Cohen’s ‘*d*’ (Cohen, 1992; Cohen, 1994), was large in the case of Neuroticism (N), Extraversion (E), and Conscientiousness (C); medium in the case of Openness (O); and small for Agreeableness (A). According to Cohen (1992), an effect size of .2 is small), .5 is medium, and .8 is large. None of the Confidence Intervals contained zero.

Table 3
Descriptive Statistics for the Full Sample of 443 Participants

	Number of Participants	Mean	Std. Deviation	Effect Size (Cohen’s <i>d</i>)	95% Confidence Interval
N	443	59.2 (79.1)	17.6 (21.2)	1.14	1.02 – 1.26
E	443	132.8 (109.4)	13.5 (18.4)	1.37	1.25 – 1.50
O	443	121.8 (110.6)	15.2 (17.3)	0.67	.56 - .79
A	443	128.0 (124.3)	14.1 (15.8)	0.24	.13 - .35
C	443	142.2 (123.6)	15.1 (17.6)	1.10	.98 – 1.22
AGE	443	38	7.9		

Notes. The means for the norming group for the NEO PI-R are shown in brackets. The standard deviations for the norming group are also shown in brackets. C – Conscientiousness, N – Neuroticism, E – Extraversion, O-Openness, A – Agreeableness. 29.8% of the sample was female.

Table 3 shows that the participants in the field study were more emotionally stable and conscientious than the general population, as evidenced by the comparisons with the norming groups in the test publisher's manual (Costa & McCrae, 1992). The means were higher and the standard deviations were lower, a result which is consistent with Salgado's (2016) research. Effect Size 'd' compares the descriptive statistics for the participants with the norm group statistics for means. The bespoke BIDR-IM mean score was 7.21 with a standard deviation of 3.73. This confirms that the field study sample was a restricted sample with respect to personality traits compared to the general population norm group for the NEO PI-R.

Table 4
Comparison of Big Five Mean Scores

	All Participants	Participants with an bespoke BIDR-IM score less than 12 - (A)	Participants with an bespoke BIDR-IM score equal to or greater than 12 - (B)	(A) and (B) Difference Effect Size (d)	95% Confidence Interval for 'd'
N	59.2	61.2	42.2	1.14	.83 - 1.45
E	132.8	132.1	138.5	0.48	.18 - .78
O	121.8	121.9	121.8	0	-.35 - .26
A	128.0	127.0	136.3	0.67	.37 - .98
C	142.2	140.3	158.4	1.29	.97 - 1.6
n	443	396	47		
	100%	(89.4%)	(10.6%)		

Note. bespoke BIDR-IM refers to scores on the Impression Management measure.
n – number of participants. C – Conscientiousness, N – Neuroticism, E – Extraversion, O – Openness, A – Agreeableness.

Table 4 contains the mean scores for the Big Five dimensions of personality for the full sample of participants – those participants who scored less than 12 on the bespoke BIDR-IM (Impression Management Measure) measure, and participants who scored 12 or higher on the bespoke BIDR-IM measure. Section 8.2 explains why a cut-off score of 12 was selected. The last two columns in the table contain the Cohen’s ‘d’ effect size comparing participants with an bespoke BIDR-IM score less than 12 with those who scored 12 or higher, and the confidence intervals for the ‘d’ scores. Table 4 shows that those with high bespoke BIDR-IM scale scores also scored higher on the Big Five dimensions than the majority of participants included in the study, with the exception of Openness. The effect size of the difference for all the dimensions were medium to large, with the exception of Openness, for those with a bespoke BIDR-IM score of 12 or higher. The confidence interval of Cohen’s ‘d’ for Openness included zero. The two groups, however, were not significantly different with respect to general cognitive ability as measured by the AH4 test - mean score for all participants was 94.4, and 89.9 for those with an IM score 12 or higher. In addition, the intercorrelations between the Big Five dimensions for the sample were broadly similar to those for the NEO PI-R norming group as Table 5 below shows.

Tables 3 and 4 show that the effect sizes were meaningful when comparing the participants with the NEO PI-R norming group contained in the test publisher manual. The effect sizes of those with a bespoke BIDR-IM score greater than or equal to 12 were also meaningful except for Openness. A comparison of the two correlation matrices in Table 5, using the Box’s M Test to test the multivariate homogeneity of variance-covariance matrices assumption, showed that χ^2 (13.9, df=15) was below the

critical value of 37.7 at $p=.001$, so the intercorrelations between the five dimensions were not statistically different from those of the norming group.

Unlike Vasilopoulos, Cucina, and McElreath (2005), who found that stronger correlations were obtained between the Big Five measures of personality and cognitive ability when a warning of verification was present, this field study found non significant and very low correlations between the AH4 cognitive ability measure in the test battery and the Big Five dimensions – N (.09), E (.03), A (.04), and C (.04). The one exception was Openness which had a statistically significant, but low, correlation of .21 ($p < .01$) with the ability measure used. The field study findings are more consistent with extant research on the relationship between the Big Five and cognitive ability (Moutafi, Furnham, & Crump, 2006) than those of Vasilopoulos et al. (2005).

Table 5
Intercorrelations between the Big Five Dimensions of the Managerial Sample and the NEO PI-R Norm Group

		Sample	N	E	O	A	C
1.	N	Study Sample	-				
		NEO Norm Group					
2.	E	Study Sample	-.25**	-			
		NEO Norm Group	-.21				
3.	O	Study Sample	-.04	.29**	-		
		NEO Norm Group	-.02	.40			
4.	A	Study Sample	-.26**	.10	.10*	-	
		NEO Norm Group	-.25	-.04	-.02		
5.	C	Study Sample	-.50**	.27**	0	.17**	-
		NEO Norm Group	-.53	.27	-.02	.24	

Notes. ** Correlation is significant at the 0.01 level (2-tailed).

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

C – Conscientiousness, N – Neuroticism, E – Extraversion, O- Openness, A – Agreeableness.

The correlations that Moutafi et al. (2006) found between cognitive ability and the Big Five, as measured by the NEO PI-R in a sample of 2658 participants, were N (.01), E (-.03), A (.01) - all of which were non significant – and O (.09) and C (-.11) both of which were significant ($p < .01$). By comparison, Vasilopoulos et al. (2006) found that the correlation for N was (-.33, $p < .01$), which was the only Big Five dimension for which a direct comparison could be made.

The results of this section show that the comparison with the norms for the NEO PI-R provides some evidence of the content relevance, representativeness, and technical quality (Messick, 1995) of the measure. This conclusion was arrived at because the results found with the Managerial sample of this research programme are consistent with the extant research on the descriptive statistics that is relevant to the NEO PI-R.

8.1.1.1. Comparison of Managerial and Validation Samples

Descriptive Statistics

Data for summed scores on the NEO PI-R Big Five facets for participants with the bespoke BIDR IM scores above 12 were not available for the Validation sample. Hence the comparisons provided in Tables 6 and 7 were for those participants with scores of less than 12 on the bespoke BIDR IM measure in both samples. The Cronbach's alpha reliabilities for the full Managerial sample respectively were Neuroticism (.77), Extraversion (.71), Openness (.71), Agreeableness (.71), and Conscientiousness (.77). The reliability figures for the Validation sample were Neuroticism (.83), Extraversion (.76), Openness (.69), Agreeableness (.74), and

Conscientiousness (.78). These reliability figures were based on the summed facet scores.

Table 6
Intercorrelations between the Big Five Dimensions of the Managerial Sample and the Validation Sample for participants with IM scores < 12

		Sample	N	E	O	A	C
1.	N	Managerial	-				
		Validation					
2.	E	Managerial	-.25**	-			
		Validation	-.29**				
3.	O	Managerial	-.04	.29**	-		
		Validation	-.14*	.29**			
4.	A	Managerial	-.26**	.10	.10*	-	
		Validation	-.39**	-.20**	-.24**		
5.	C	Managerial	-.50**	.27**	0	.17**	-
		Validation	-.54**	.39**	-.10	.28**	

Notes. ** Correlation is significant at the 0.01 level (2-tailed).

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

C – Conscientiousness, N – Neuroticism, E – Extraversion, O- Openness, A – Agreeableness.

The next table contains details of the means and standard deviations for the two field study samples. The data provided is a comparison between the Managerial and Validation samples for those participants with a bespoke BIDR IM scores less than 12.

Table 7
Means and Standard Deviations for the Big Managerial and the Validation Samples

	Sample	N	E	O	A	C
Mean	Managerial	60.9	127.2	111.3	131.1	141.4
	Validation	61.2	132.1	121.9	127.0	140.3
Standard Deviation	Managerial	16.1	14.1	15.7	13.9	12.5
	Validation	17.0	13.2	15.2	13.7	14.4

Notes. C – Conscientiousness, N – Neuroticism, E – Extraversion, O-Openness, A – Agreeableness. Managerial Sample Size n=396. Validation Sample Size n=201.

8.1.2 Exploratory Factor Analysis (EFA)

The results of the Exploratory Factor Analysis (EFA) for two sets of participants – the full set of 443 cases and the reduced set of 396, i.e. those with a score less than 12 on the bespoke BIDR-IM measure are shown in Table 8. Two higher order factors were extracted with each having an eigenvalue greater than one using Principal Component Analysis; this factor analytic method was used by Costa and McCrae (1992) in the development of the NEO, and Direct Oblimin rotation. This factor structure was as expected based on previous published studies (Costa & McCrae, 1995) and provides some substantive (Messick, 1995) evidence for construct validity. The variance accounted by the first EFA factor extracted is sometimes used as a measure of how strongly personality measures are saturated with a putative

General Factor of Personality, or GFP, (Davies, Connelly, Ones, & Birkland, 2016; Musek, 2007). However, as mentioned previously in Chapter 4, Lance and Jackson (2015) have questioned the validity of this approach in general when it comes to identifying putative general factors.

Table 8
Variance Accounted for by Factors with an eigenvalue > 1

	All Participants	Participants with BIDR-IM <12
First Factor	39.55%	37.24%
Second Factor	22.97%	22.82%
Cumulative Variance	62.52%	60.06%
Factor Correlation	.18	.19

Table 9 shows the factor loadings on the two higher order factors, with factor loadings greater than .4 highlighted in bold.

Table 9
EFA Factor Loadings on the Two Higher Order Factors

Big Five Dimension	All Participants		Participants with BIDR-IM <12	
	Factor 1	Factor 2	Factor 1	Factor 2
E	.324	.656	.294	.660
O	-.180	.908	-.171	.905
A	.540	.048	.474	.081
C	.842	-.058	.830	-.075
N	-.852	.025	-.844	.042

Notes. Factor loadings greater than .4 are in bold print. C – Conscientiousness, N – Neuroticism, E – Extraversion, O-Openness, A – Agreeableness.

These results are consistent with the findings of other research that two higher order factors explain more of the variance than a single higher order factor (Digman, 1997). The pattern of factor loadings was as expected based on previous research (Costa & McCrae, 1995; Digman, 1997). Specifically, Conscientiousness, Agreeableness and Emotional Stability (Neuroticism reversed scored) loaded on one factor, and Extraversion and Openness loaded on the other higher order factor. The exclusion of those who scored high on impression management did not alter the substantive structural findings from the EFA analysis. The two factors extracted in the analysis were correlated, but were found to have a relatively low correlation.

The next two tables contained a comparison of the EFA analysis for the two field study samples. Once again the comparisons are for those participants scoring less than 12 on the bespoke BIDR IM measure. Table 10 provides details of the total variance accounted for by each factor extracted using Principal Component Analysis with Direct Oblimin rotation.

Table 10
Variance Accounted for by Factors with an Eigenvalue > 1 in the Validation and Managerial Samples for Participants with an IM score <12

	Validation Sample	Managerial Sample
First Factor	41.76%	37.24%
Second Factor	23.89%	22.82%
Cumulative Variance	65.65%	60.06%
Factor Correlation	.19	.18

The next table, Table 11, contains a comparison of the factor loadings on the factors extracted for both samples.

Table 11

EFA Factor Loadings on the Two Higher Order Factors in the Validation and Managerial Samples for Participants with an IM score <12

Big Five Dimension	Validation Sample		Managerial Sample	
	Factor 1	Factor 2	Factor 1	Factor 2
E	.495	.446	.294	.660
O	-.135	.944	-.171	.905
A	.474	.354	.474	.081
C	.821	.014	.830	-.075
N	-.901	.305	-.844	.042

Notes. Factor loadings greater than .4 are in bold print.

C- Conscientiousness, N – Neuroticism, E – Extraversion, O-Openness, A – Agreeableness.

The ratio of subjects to items in an EFA has a significant and substantial influence on factor loadings matrix (Osborne and Costello, 2004). The odds of getting a correct factor pattern matrix increases as the sample size increases. This is also true as the ratio of subjects to items increases (Osborne, 2014). The loadings are the regression coefficients of the factor model equation and even with equal sized validation samples, the regression coefficients are not stable (Neter, Kutner, Nachtsheim, & Wasserman, 1996). It is, for this reason, that unit weighted linear models outperform regression models (Einhorn & Hogarth, 1975; Wainer, 1976). In addition, the factor loading pattern matrix is not very stable from sample to sample (Osborne, 2014). Finally, according to Hensen and Roberts (2006), CFA is a more

appropriate method of analysis when there are a priori expectations concerning the factor structure of a measure. This is likely to be true when the instrument is not new and when there is extensive knowledge of the factor structure of the measure.

8.1.3 Confirmatory Factor Analysis (CFA)

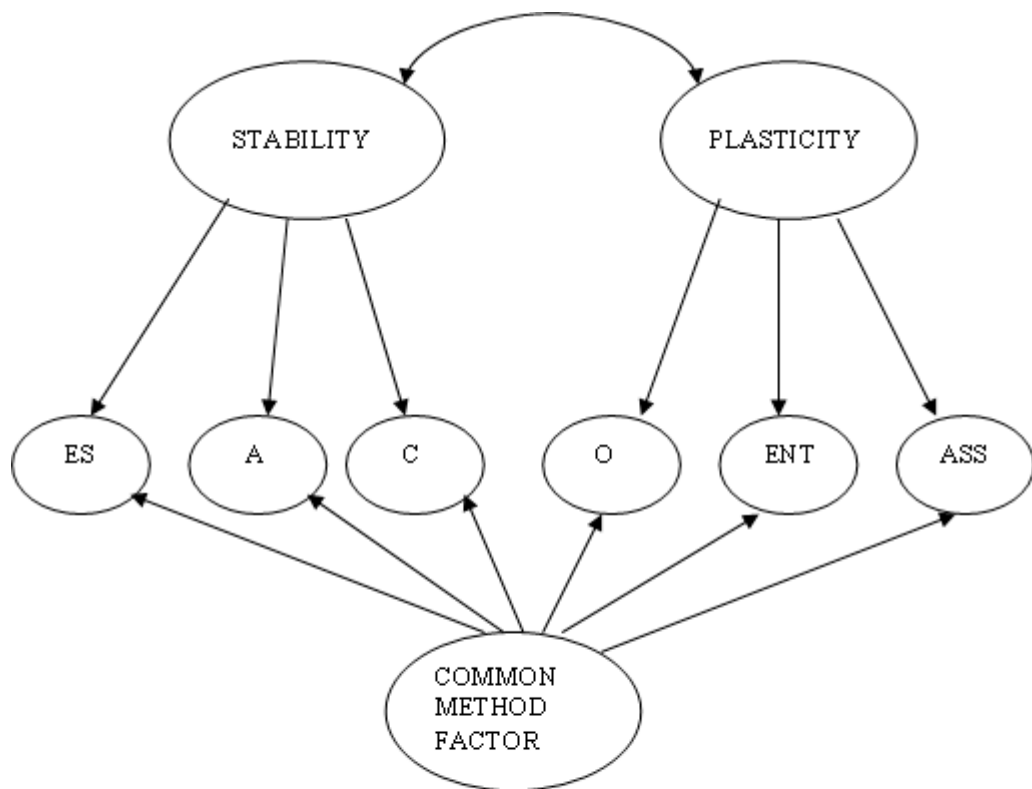
A number of different Confirmatory Factor Analysis (CFA) models were tested in order to evaluate the internal structural aspect of strong construct validity (Messick, 1995), i.e. “the fidelity of the scoring structure to the structure of the construct domain at issue” (p. 745). This includes CFA model testing of the Big Five with a GFP and, alternatively, the Big Five with two higher order factors. Because of improper solutions and identification problems encountered with these models some additional models were also tested. As will be seen there were major identification problems that arose in the model evaluations that required a detailed and on-going technical problem resolution approach to be taken as the results appeared. This resolution inevitably relied very heavily on the technical discourse, detailed in Section 6.2 of the previous chapter, concerning the issues that arose during the analyses. Table 12 contains a summary of the details of the models tested together with goodness of fit statistics and factor loadings.

Table 12
CFA Goodness of Fit Indices for Participants with Impression Management scores less than 12

Model Tested	χ^2	df	SMRM	RMSEA**	CFI	TLI	Comment
1. 5 Indicators with a single GFP	42.5 (p<.01)	5	.0759	.139	.824	.647	Loadings E .35, A .29, O .08, ES .72, C .73. The Loading of O was not significant.
2. 5 Indicators with a single GFP, and MI's	6.3 (p=.177)	4	.0283	.039 (0 - .09)	.989	.973	The errors of E and O were allowed to correlate. Loadings similar to Model 1.
3. 5 Indicators with 2 Higher Order Factors	Inadmissible Solution	4	One parameter with negative variance, two parameters with very large SE's, one indicator had a Squared Multiple Correlation > 1				
4. Model 3 with Equal Higher Order Loadings for Plasticity	Inadmissible Solution	2	Three parameters were found to have negative variance				
5. 6 Indicators with a single GFP	175.7 (p<.01)	9	.1151	.219	.570	.284	Loadings ENT .37, ASS .48, O .13, A .25, ES .67, C .70
6. 6 Indicators with 2 Higher Order Factors	106.5 (p<.01)	8	.0803	.178	.746	.524	Loadings ENT .61, ASS .71, O .33, A .27, ES .73, C .76
7. Model 6 and a Common Method Factor	Inadmissible Solution	7	One parameter with a negative variance. There was a non positive defined matrix for Stability and Plasticity				
8. 6 Indicators, 2 Higher Order Factors, and MI's	36.7 (p<.01)	7	.0619	.105	.923	.836	ASS error was allowed to correlate with A error. Loadings were similar to Model 6.
9. Model 8 and a Common Method Factor	Inadmissible Solution	6	Non positive definite matrix between the residuals of Openness (O) and Emotional Stability (ES)				

Notes. χ^2 – chi squared statistic; df – degrees of freedom; TLI - Tucker–Lewis index; CFI - comparative fit index; RMSEA - root mean square error of approximation. SMRM - standardised root mean square residual; MI – Modification Indices. ** 90% Confidence Interval for REMSA is only shown when the value is suggestive of acceptable model fit.

All models tested using CFA, and reported in Table 12, were based on the sample of participants with bespoke BIDR-IM scores less than 12. Outliers identified by the Mahalanobis D^2 criterion (Kline, 2011) were also removed from the analysis, reducing the sample size with a bespoke BIDR-IM score less than 12 to 388. The sample was also examined for kurtosis, skewness and multivariate normality, none of which were found to be an issue. The models evaluated were all variants of the model depicted in Figure 8.



Notes. ES – Neuroticism, reversed scored, A – Agreeableness, C – Conscientiousness, O – Openness, ENT – Enthusiasm, ASS – Assertiveness

Figure 8 Illustration of Model 7 tested in Confirmatory Factor Analysis

The guidelines suggested by Brown (2006) were followed in deciding which fit measures to use, as well as the recommendations of Hu and Bentler (1999) and others (McDonald & Ho, 2002), summarised in Hooper et al. (2008). SRMR values should be not greater than .08; RMSEA values should not be greater than .06/.07; a value of CFI = or > .95 is indicative of good fit; TLI values should have a threshold of = or > .95. These fit indices, taken together, can be suggestive of an acceptable model. The relationship of the models tested to the substantive theory of personality was also included as a necessary consideration in the model evaluations when evaluating the suggested modification indices (Byrne, 2010; Hooper, Coughlan, & Mullen 2008; Kline, 2011). This was evaluated by an examination of the factor loadings of the five broad dimensions on the higher order latent factors, and the variance accounted for by the indicators of the five dimensions. In addition, a close examination of the standardised residuals so as to look for localised ‘strain’ in the model tested (Byrne, 2010; Kline, 2011; McDonald, 1999).

8.1.3.1 Comments on CFA Models Tested

The models were tested in the following order. Firstly, a GFP model with the Big Five as indicators. This was then followed by testing the model with Stability and Plasticity as the two higher order factors also with five indicators. Following the problems encountered with these models the succeeding models tested were designed to overcome the poor fit or empirical underidentification issues that arose.

The first model fit was very poor – $\chi^2 = 42.5$ (df=5, $p = .177$); REMSA .139; CFI .824; TLI .647. Johnson, Rosen, and Djurdjevic (2011b) emphasise the need to

examine factor loadings as well as fit statistics. They make the point that “Therefore, organizational scholars should not focus entirely on changes in fit statistics when considering the effects of CMV but instead should consider how factor loadings, path coefficients, and effect sizes are influenced by implementing various controls of CMV” (p. 759). The loadings found were as follows for the five factor indicators loading on the putative GFP - Emotional Stability .72, Extraversion .35, Openness .08, Agreeableness .29, and Conscientiousness .73. The factor loading for Openness was very poor and not significant, and the loading for Extraversion and Agreeableness were both poor. These loading results are broadly similar to those in Chang et al. (2012)’s meta-analysis - Emotional Stability .57, Extraversion .03, Openness .09, Agreeableness .48, and Conscientiousness .38. There were no significant differences between the model tested with the full sample of 443 participants or the sample of 388 with an IM score less than 12. The errors of Extraversion and Openness were allowed to correlate in Model 2, as suggested by the modification indices (MI’s) of Model 1. This approach is consistent with the recommendation of Landis, Edwards, and Cortina (2009) that correlated errors should only be allowed when a strong a priori reason exists. This was the only MI used in the analysis of this model because of the theoretical relationship between Extraversion and Openness due to primary and secondary factor loadings, as described in Chapter 3.

Allowing the residual errors to correlate in Model 2 did improve the model fit. As expected, there was no change of any significance in the factor loadings, which in the case of Extraversion, Agreeableness, and Openness accounted for only a very small proportion of the total variance in each individual case. An examination of the AMOS CFA results showed that the standardised residual covariance estimate between Openness and Agreeableness was borderline. Taking a lenient view in the

evaluation of Models 1 and 2 they provided support, if any, for a four-indicator model rather than a Big Five model. However, it is important to note that taking the factor loadings of Extraversion, Agreeableness and Openness into account the variance accounted by each of these three Big Five dimensions was very low.

Models 3 and 4 were tests of a five-indicator model with the two higher order factors, i.e. Plasticity and Stability. It is of interest here that Kline (2011, p. 238) makes the very pertinent point that if the two higher order factors correlation is set to one in the CFA analysis then Model 3 becomes equivalent to Model 1. The one factor Model 1 is, in fact, a restricted version of Model 3. Using this reasoning Model 3 is a test of Model 1 with, de facto, the correlation between Stability and Plasticity freed up. The result for Model 3 was inadmissible. This model was also tested (Model 4) using equal loadings for Extraversion and Openness loading on Plasticity, as recommended by Kenny, Kashy, and Bolger (1998) and Kline (2011) for a two higher order factor model with only two indicators loading on one of the factors in the model. However, neither did this model yield an admissible solution.

Because of the failure to obtain a solution for the two higher order models tested in Models 3 and 4 it was not possible to directly compare the Stability/Plasticity higher order structure with the GFP theory of the higher order structure and therefore to carry out a simple comparison test of the hypotheses of Chapter 7. Using the reasoning of Kline (2011, p. 238) it can be argued that one putative reason for this failure to find an admissible solution for Model 3 is that the two factor, five-indicator model was empirically underidentified due to the two higher order factor not being correlated (Brown, 2006; Kline, 2011; McDonald, 1999). To test if the empirical underidentification was due to the low loading of Openness on Plasticity a two higher order model was also tested that omitted Openness as an indicator for Plasticity. The

results are not shown in Table 12 because this model was also empirically unidentified suggesting that the identification problem is due to no Stability/Plasticity correlation.

An additional model was tested which was based on the recent Davies, Connelly, Ones, and Birkland (2016) meta-analytic study using a bifactor model. In addition to the two uncorrelated Stability/Plasticity higher order factors, this model also contained a general factor loading on the five indicators. This model is also not shown, unlike the findings of Davies et al. (2015), because it too was empirically unidentified.

In order to allow the competing theories of the higher order structure of the Big Five to be further tested against each other in this research programme, the NEO PI-R Extraversion facets of Warmth, Gregariousness and Positive Emotions were assigned to a new indicator 'Enthusiasm', and the other three facets – Assertiveness, Activity and Excitement Seeking - were assigned to a second new indicator 'Assertiveness', along the lines suggested by the findings of aspects between the facets and the Big Five dimensions described in Chapter 3 of DeYoung, Quilty, and Peterson (2007). A similar approach was taken by Roberts, Walton and Viechtbauer (2006) in their meta-analysis of the stability of the Big Five over lifetime in which they, too, partitioned Extraversion into two aspects of Social Vitality and Social Dominance. By following this methodological strategy, it was possible for the putative Plasticity factor to have three indicators. If Stability and Plasticity are not substantively correlated this model will then meet the minimum requirement of 'just identified' for model empirical identification (Kline, 2011)

A single GFP model was tested using the six indicators (Model 5). This model showed very poor fit, with some weak loadings, which was only marginally improved by allowing the errors of Assertiveness and Agreeableness, as suggested by the MI's, to correlate. The next defensible MI change, suggested by the AMOS analysis, added to Model 5 which was not included in Table 12 because it also resulted in an inadmissible solution.

Following this Model 6, a two-factor model with three indicators loading on each of Plasticity and Stability was tested. The model was an improvement on Model 5 but the fit of this model was poor. Then in Model 7, this same two factor model using the Unmeasured Latent Method Factor Technique (Podsakoff et al., 2003) with a common method factor loading on the six indicators was tested (Model 7 and Figure 8). The common method factor was constrained to load equally on the six indicators. This model would also be described as a 'bifactor' model (Davies et al., 2016; McDonald, 1999) if there was no correlation between the factors.

Using the procedure followed by Kline (2011, p. 238) Model 6 was compared with Model 5, i.e. the six indicators with two higher order factors model with the GFP model with 6 indicators. The Chi Squared Difference between the two models was above the critical value ($\Delta\chi^2=69.4$, $df=1$, $\Delta\chi^2_{crit}=7.88$). The goodness of fit measures were poor for both models although most of the factor loadings were stronger for Model 6 compared with Model 5 – ES (.27 v .67), A (.7 v .25), C (.76 v .7), Enthusiasm (.61 v .37), Assertiveness (.71 v .48), and O (.33 v .13). Next, a two-factor model with the errors of Agreeableness and Assertiveness allowed to correlate, as suggested by the modification indices (Model 8), and was tested. This MI was consistent with evidence for NEO PI-R item primary and secondary factor

loadings (Johnson, 1994) covered in Chapters 3 and 6. The fit for this model was poor. Finally, a variation of Model 8 was also tested using a common method factor model with equal loadings on the six indicators (Model 9). Models 8 and 9 comply with the Kenny, Kashy and Bolger heuristics (Brown, 2006) for non-standard CFA models with correlated errors. Model 8 was the second ‘best’ model of the nine CFA models tested (the ‘best’ being Model 2 based on fit indices only). However, only one of the fit indices, SMRM, for this model met the suggested cut-off values of Hu and Bentler (1999). In addition, an examination of the standardised residual covariances showed that there were three very poor covariance estimates - O and ENT, O and C, and A and ENT. McDonald (1999) places strong emphasis on this aspect of model fit evaluation in deciding whether or not the CFA model being tested is adequate.

In addition, Kline (2011) advises researchers to consider the value of χ^2 as well when the fit indices are poor or marginal. For this reason, the analyses of Models 1 and 3 were relied upon in determining whether Stability and Plasticity were correlated in this research programme. Arguably they are not on the basis that the only difference between Models 1 and 3 was that the single higher order GFP was replaced by two higher order factors Stability and Plasticity. It is arguably a plausible reason that an inadmissible solution was found for Model 3 due to empirical under-identification arising from the inference that these two higher order factors were not correlated. Both Models 1 and 3 had the exact same indicators and the analysis used the same raw data.

A further check on the accuracy CFA results obtained in Table 12 was carried out by comparing the CFA results using the Big Five dimension ‘summed scores’ with the CFA results based on a hierarchical CFA models with the six facets loading on each Big Five dimension. The results obtained are shown in Table 13. The data

used for this part of the analysis was from the Validation field study data for the practical reason that all facet scores were readily available from this field study, so the figures for the various indices are different from those in Table 12. The ‘summed score’ model is shown first in each case using the model numbering system used in Table 12. The ‘facet score’ models have the letter ‘A’ appended to the model number of Table 12.

For each of the models, the Mardia Coefficient of multivariate normality was checked. According to Bentler (1995), “In practice, values larger than 3 provide evidence of nontrivial positive kurtosis, though modelling values may not be affected until values are 5, 6, or beyond” (p. 106). All of the models based on ‘summed scores’ met the Mardia Coefficient recommended values. However, for the ‘facet score’ models the Mardia Coefficients were 14 or higher. These high Coefficients can easily lead to estimation problems with algorithms using maximum likelihood (ML) estimation algorithms (Kline, 2011). Because of multivariate normality concerns in ML estimation, the ‘summed scores’ CFA model results were relied upon in evaluating the outcomes of this research programme because Brown (2006) recommends against using ML when non-normality is excessive.

Table 13
Comparison of 'Summed Score' CFA Models with 'Facet Score' Models

Model Tested	χ^2	df	SMRM	RMSEA	CFI	TLI	Comment
1. Five Factors and GFP	52.6 (p<.01)	5	.1015	.218	.721	.442	The loading of O was not significant. Loadings E .47, A .42, O .11, ES .67, C .78
1A. Five Factors with a GFP	1252.4	400	.1246	.103	.641	.610	The loading of O was not significant. Loadings E .68, A .62, O .10, ES .78, C .85
3. 5 Factors - 2 Higher Order Factors	Inadmissible Solution	4	Heywood case – the disturbance of Extraversion had a negative value				
3A. 5 Factors with two Higher Order Factors	Inadmissible Solution	399	Heywood case – the residual of Extraversion had a negative variance				
5. Six Factors and GFP	133.1 (p<.01)	9	.1288	.263	.572	.287	Loadings ENT .84, ASS .63, O .32, A .39, ES .40, C .49
5A. Six Factors and GFP	1272.2	404	.1326	.104	.663	.601	Loadings ENT .95, ASS .6, O .32, A .47, ES .71, C .72
6. 6 Factors – 2 Higher Order Factors	79.3 (p<.01)	8	.0908	.211	.746	.524	Loadings ENT .61, ASS .71, O .33, A .7, ES .27, C .76
6A. Six Factors – 2 Higher Order Factors	Inadmissible Solution	Heywood case – the residual of Extraversion had a negative variance					

Notes. χ^2 – chi squared statistic; df – degrees of freedom; TLI - Tucker-Lewis index; CFI - comparative fit index; RMSEA - root mean square error of approximation. SMRM - standardised root mean square residual.

Because of the failure to find a proper model or the poor model fit of those models with a proper solution one further test for the absence or presence of a correlation between Stability and Plasticity was carried out. For this test both Stability and Plasticity were treated as stand alone 'just identified' (see Chapter 6, Section 6.2) CFA models and subjected to a CFA evaluation, using AMOS. The factor loadings for these two stand alone models were then compared with the factor loadings of

Model 6, i.e. the CFA model with three indicators loading on the two putative higher order factors that were allowed to correlate.

8.1.3.2 Just Identified Model Comparison

In carrying out this analysis the following analytical procedure was adopted. Both higher order factors were treated as separate factors and separately subjected to a CFA. The CFA for both of these factor models with three indicators each are ‘just identified’ CFA models, each with a unique solution (Brown, 2006). The predicted covariances and variances of any congeneric indicators’ factor loadings on each factor in a CFA are determined by the maximum likelihood estimated factor variance, factor loadings and error variances (Brown, 2006; Kline, 2011). This is also true even if the indicators of a factor cross load on another factor because of a covariance between the two factors (Brown, 2006). The loadings control the degree to which the indicators of the same factor are calculated to be related to each other. They also control how much of the variance in the indicator is due to error. Indicators from different factors are only related if the factors are correlated. If the two higher order factors of Plasticity and Stability are not correlated then mathematically the factor loadings should be similar in size when the two factors with their respective indicators are together subjected to a CFA in the same model, as well as when each factor is separately subjected to a CFA with its three indicators, if the CFA of the two factors analysed together results in a good fit for the combined model. It can, therefore, be plausibly argued that in the case of an unacceptable model, such as Model 6, with the same factor loadings as the two ‘just identified’ models the lack of fit is due to problems in the model arising from a low, or no, correlation between the two factors.

The reason for this is that the estimated population variance/covariance matrix for each model in a CFA is constructed from the estimated factor loadings, factor variances, and factor covariances of the algorithm (Brown, 2006; Kline, 2011, McDonald, 1999). If the benchmark congeneric ‘just identified’ factor loadings are also found to be the same in an unacceptable (from a fit perspective) higher order factor model e.g. Model 6 in Table 12, the cause of the poor fit is arguably due to an incorrect factor correlation, between the two factors F_1 and F_2 , estimate. This is because estimated model covariances are calculated by multiplying the product of the two estimated unstandardised factor loadings. For example, say λ_{11} and λ_{21} are two of three indicators loading on F_1 , and λ_{42} is one of three indicators loading on F_2 . The predicted factor covariances in a two higher order factor model equals $\lambda_{11}(\text{var } F_1) \lambda_{21}$, $\lambda_{11}(\text{covar } F_1 F_2) \lambda_{42}$, and $\lambda_{21}(\text{covar } F_1 F_2) \lambda_{42}$. In this case of a comparison of Model 6 with the ‘just identified’ model the lack of fit may be attributable to a problem with the estimation of the term $(\text{covar } F_1 F_2)$ i.e. the covariance between the two factors. An explanation for the inference that the lack of fit is due to $(\text{covar } F_1 F_2)$ can be readily seen from an examination of Figure 9 on the next page, which was taken directly from Hoffman (2017). In a just identified model of two uncorrelated higher order factors, each with three indicators, only the variances and covariances within the Blue boxes of Figure 9 are estimated. If the two factors are correlated then there are other covariances, outside the Blue boxes, which are due to cross loadings of the indicators on both factors. These covariances are estimated with the same factor loadings that are found within the Blue boxes in Figure 9. Thus the error in the model fitting is due to an error in the estimation of the covariance between F_1 and F_2 .

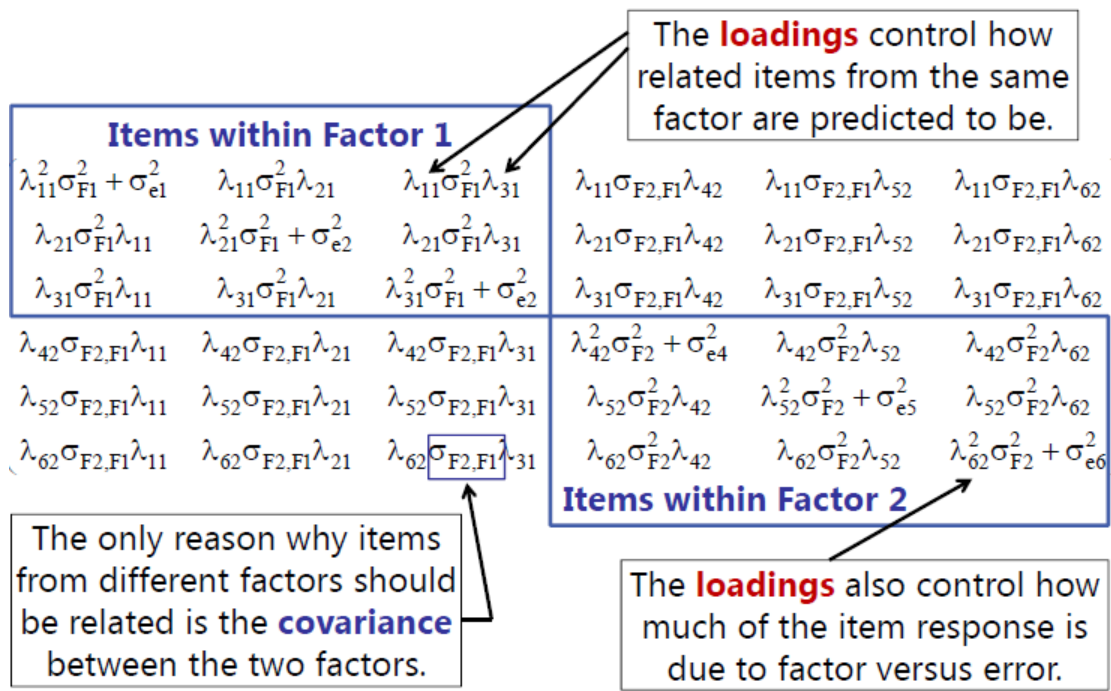


Figure 9 Two Higher Order Factor Model – Model Predicted Item Covariance Matrix (Hoffman, 2017)

A chi squared test was carried out using the loadings of the three indicators for each of the two just identified factor model (the expected loadings from a model with two uncorrelated higher order factors) as the expected model. This was compared with a model using the estimated factor loadings for Model 5 in Table 12 as the observed factor loadings and ignoring the factor loadings due to cross correlations e.g. $\lambda_1(\text{covar } F_1F_2)\lambda_4$ in the preceding paragraph. The same goodness of fit exercise was performed on the Validation sample. The absence of a significant difference between the expected and observed factor loadings was taken as evidence of no correlation between Stability and Plasticity. The results of this analytical procedure are shown in Table 14 below for both Managerial and Validation field study samples.

Table 14
Comparison of expected and observed factor loadings

Indicator	<u>MANAGERS</u>		<u>VALIDATION SAMPLE</u>	
	Two CFA Just Identified Models Combined	CFA Model 6 with two Higher Order Factors	Two CFA Just Identified Models Combined	CFA Model 6 with two Higher Order Factors
C	.692	.757	.689	.757
A	.287	.270	.331	.322
ES	.755	.695	.763	.696
ENT	.775	.612	.722	.687
ASS	.544	.711	.639	.694
O	.381	.328	.405	.365
χ^2 (11.05 crit*)	.051		.012	

Note. crit*- statistically significant at .05 level

The χ^2 statistics were not significant for either sample since they were less than the critical value of χ^2 for five degrees of freedom. This consistency between the ‘just identified’ models’ factor loadings and the equivalent Model 6 loadings suggest that the very poor fit of Model 6 is mainly due to inaccurate estimates of the covariance between the two different higher order factors.

8.1.3.3 Invariance Analysis

Generalisability is one of the aspects necessary for establishing the construct validity of a measure (Messick, 1995; McDonald, 1999; Vandenberg & Lance, 2000;

see also Chapter 2 of this thesis). Does the measure used in one applied or research situation behave in a similar manner when used with different groups of participants?

As described in Chapter 7 the Validation sample consisted of job applicants for a high level technical job in a single organisation. The validation procedure, based Vandenberg and Lance (2000), was an integral part a strong construct validation exercise that incorporates generalisability (Messick, 1995), because mean scores for the two groups on the Big Five differed on some of the personality dimensions. For this exercise CFA Model 3 from Table 12 was used. Even though the solutions found for the analyses of Table 12 were not admissible the objective of the invariance test was to determine if the model operated in an invariant manner in both field study samples. The results of the procedure described in detail in Chapter 7 are presented in Table 15.

Table 15
Invariance tests of Managerial and Validity samples

Model	χ^2	SRMR	REMSA	CFI	$\Delta\chi^2$	ΔCFI
Configural	33.21	.0283	.073	.936	-	-
Equal Loadings	42.02	.0351	.065	.926	8.82 (p=.07)	.01
Equal Loadings and Equal Error variances	52.23	.0382	.057	.913	10.21 (p=.12)	.013
Equal Loadings, Equal Error variances and Equal Factor variances and covariances	53.11	.0352	.055	.913	.88 (p=.35)	-
Equal factor loadings and intercepts	116.9	.0344	.100	.746	74.86* (p < .01)	.177

Note. * Statistically significant. $\Delta\chi^2$ – change in chi squared statistic; df – degrees of freedom; TLI - Tucker–Lewis index; CFI - comparative fit index; RMSEA - root mean square error of approximation. SMRM - standardised root mean square residual. ΔCFI – change in CFI

The fit indices for the configural model tested were acceptable, even though the CFI is just under .95. Byrne (2010) recommends using both $\Delta\chi^2$ and ΔCFI tests of measurement invariance between the models tested in a multigroup invariance test. Unlike the assessment of model fit in CFA a $\Delta\chi^2$ is statistically significant in multigroup invariance testing when the probability is less than 5%. A threshold figure for ΔCFI of .01 or less is recommended for multigroup equality (Brown, 2010).

Table 16
Big Five Mean Scores Comparison between Field Study Samples

Big Five Dimension	Cohen's 'd'	95% Confidence Interval
Extraversion	.37	.20 to .54
Openness	.68	.50 to .85
Agreeableness	-.31	-.48 to -.14
Conscientiousness	-.07	-.24 to .11
Emotional Stability	.02	-.15 to .19

The Equal Loadings model supported group invariance in that neither $\Delta\chi^2$ (not statistically significant) nor ΔCFI was a problem. Adding an error variance equality constraint was also deemed to be acceptable. Constraining all estimated factor loadings, factor variances and covariances to be equal did not improve the fit between the measurement models of the Managers and Validation samples. These

parameters are acting in a similar manner in both samples. This confirms the initial finding from the Unconstrained model analysis of the two field samples i.e. that they were invariant. The last test of invariance in Table 15 shows that there is a difference between the two field study samples when intercepts are included. This was to be expected because of the differences in mean scores on some of the Big Five dimensions shown in Table 16, so therefore the of the statistically significant difference in intercepts in the last invariance comparison of Table 15 is consistent with the effect size differences between the two field study samples.

8.2 Cluster Analysis

To help establish the construct validity, from an empirical perspective, of the bespoke BIDR-IM measure a Cluster Analysis was used to further explore groupings among the Managerial field study. This was done using the meta-analytic study findings of Connelly and Chang (2012), as described in Chapter 6, which showed that socially desirable responding was linked to scores on the BIDR as well as an individual's standing on the Big Five dimensions of Agreeableness and Conscientiousness. A two-step cluster analysis, using SPSS Version 21, of the Managerial field study data produced a meaningful two clusters solution with a silhouette measure of cohesion and separation approaching .5. SPSS categorises any clustering that has a silhouette measure of cohesion and separation of .5 or higher as 'good'. Table 17 contains the descriptive statistics, including the mean scores, for the variables that were used to identify the two cluster groups that emerged from the analysis. There were three outliers identified by SPSS in the analysis. The most

important clustering variable was Conscientiousness followed by the bespoke BIDR-IM scale, and the least important but still meaningful clustering variable was Agreeableness.

Table 17
Mean Scores for the two Cluster Groups from clustering based on IM, Agreeableness, and Conscientiousness

Cluster	1	2
Participants (n)	271	169
Percent	61.6 %	38.4 %
Conscientiousness	149.6	125.2
Agreeableness	132.4	118.1
IM	8.95	4.47

Figures 9 and 10 contain plots of the frequency of the occurrence of cases in a particular Cluster group against IM score, for Clusters 1 and 2, superimposed on a plot for the total sample of participants. In the two figures the descriptor ‘overall’ refers to the full sample of participants. Clusters 1 and 2 are designated as ‘1’ in Figure 9, and designated ‘2’ in Figure 10.

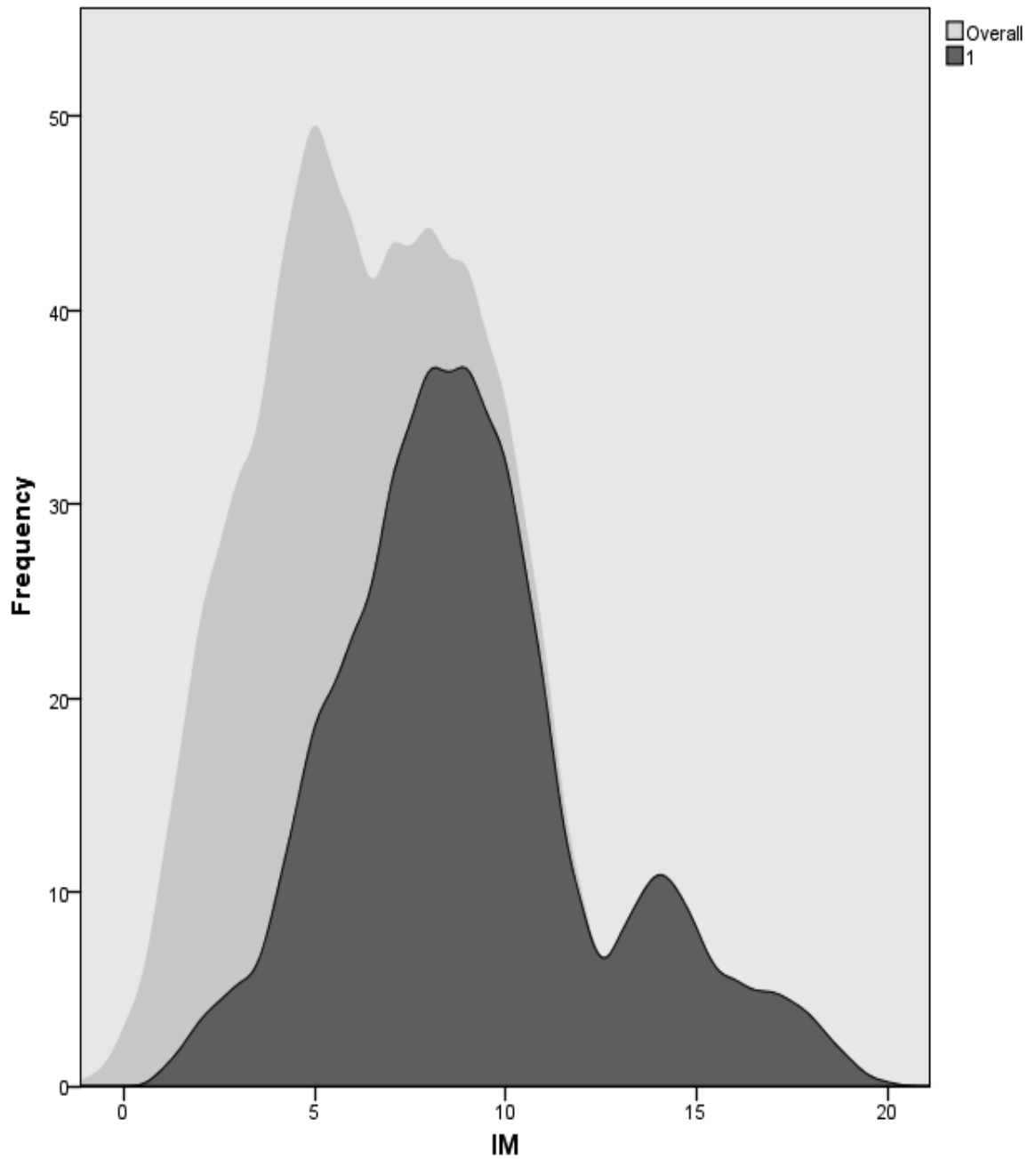


Figure 10 Plot of frequency of occurrence against IM score for Cluster 1 compared to the full sample of participants

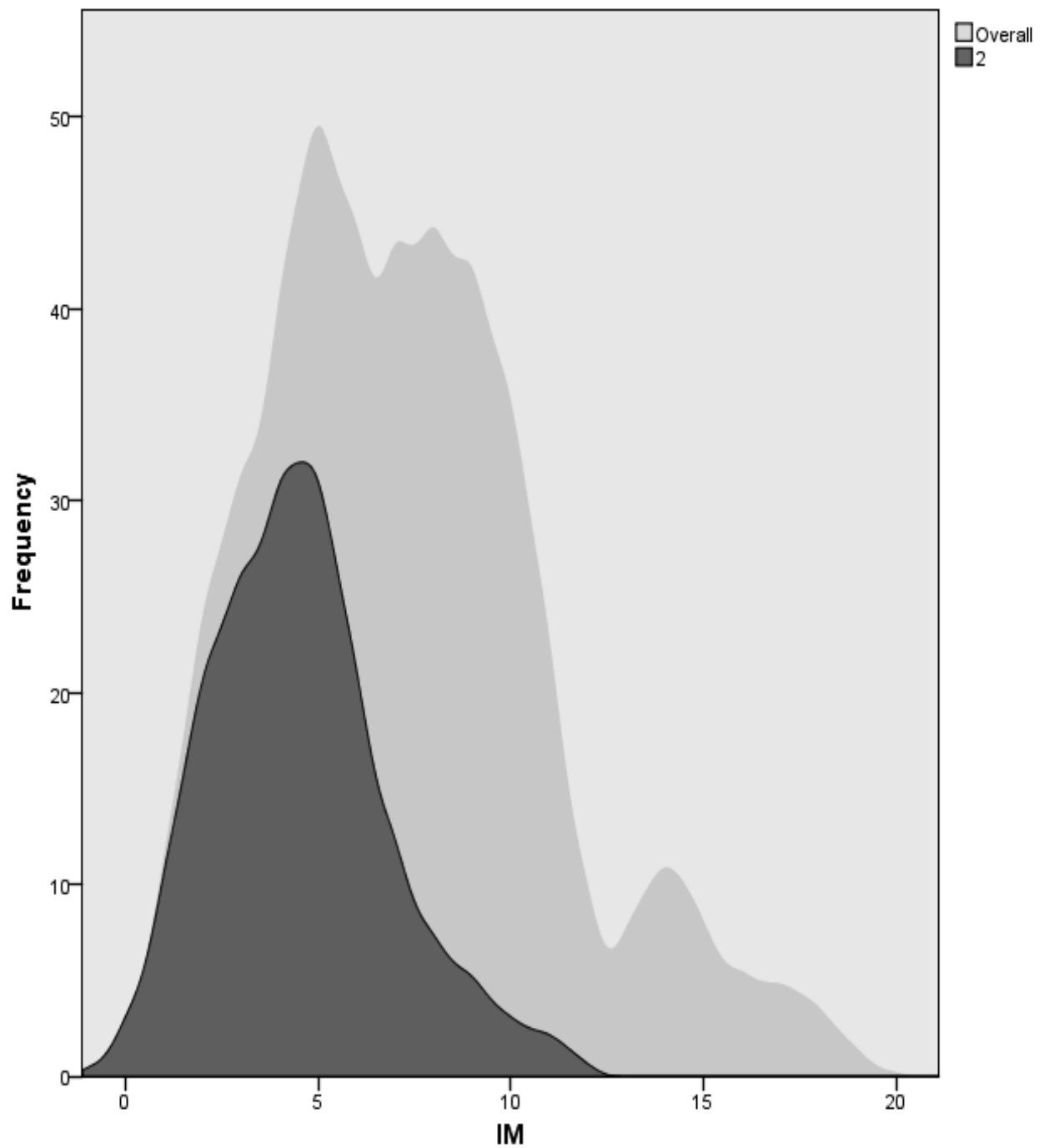


Figure 11 Plot of frequency of occurrence against IM score for Cluster 2 compared to the full sample of participants

The frequency of the number of participants who scored at each bespoke BIDR-IM score is shown in the Table 18. The table contains the frequency number for the total sample, and the frequency number for Cluster 2.

Table 18
Frequency Table of IM scores in Total and in Cluster 2

IM Score	Total Number	Number in Cluster 2
0	2	2
1	9	9
2	27	23
3	31	26
4	38	31
5	57	35
6	39	19
7	44	12
8	44	6
9	43	6
10	36	2
11	24	3
12	6	-
13	6	-
14	13	-
15	7	-
16	5	-
17	5	-
18	4	-
19	1	-
20	-	-

Note – Average scores in Cluster 2 for IM, Agreeableness, and Conscientiousness are lower than in Cluster 1

Participants who scored 12 or higher on the bespoke BIDR-IM scale were all in Cluster 1, the cluster with high IM, Agreeableness and Conscientiousness scores. Participants with an IM score of 12 or higher were grouped together and their average scores on IM, Agreeableness and Conscientiousness were compared with all other

participants, excluding outliers. Effect sizes and confidence intervals for the comparisons are shown in the next table.

Table 19
Cohen's 'd' Effect Sizes for Participants in Cluster 1 with IM Score of 12 or greater, compared with all other Participants.

	Effect Size 'd'	95% CI's
IM	3.2	2.8 to 3.5
Conscientiousness	1.3	1.0 to 1.6
Agreeableness	.7	.4 to 1.0

The effect size of the IM measure for those participants with a score of 12 or higher is extremely large, as would be expected, while the confidence intervals for Agreeableness and Conscientiousness are not as large. A large effect size would occur for IM, in any event, but what is important is the scale of the effect size for IM compared with the other two clustering variables when viewed from the perspective of the Borghans, Duckworth, Heckman, and Ter Weel (2008) equation of Chapter 6. The correlations of Conscientiousness and Agreeableness with IM are low (.2 and .09) and not significant for the sample with IM scores of 12 or higher, while Conscientiousness and Agreeableness were significantly correlated (.31). IM scores were significantly correlated with Conscientiousness (.31, $p < .01$) and Agreeableness (.28, $p < .01$) in the sample with IM scores less than 12, similar to what Paulhus (2002) found, and Conscientiousness and Agreeableness were also significantly correlated (.19, $p < .01$). However, care must be exercised in evaluating these results because of

the small sample size of those with an IM score of 12 or higher, and the likelihood that the distributions of scores are not normal for those scoring 12 or higher.

8.3 Monte Carlo Simulations

As described in detail in Chapter 7 a Monte Carlo simulation approach, using data from the Managerial field study sample, was employed in order to evaluate the consequential aspect (Embretson, 2007; Messick, 1995) of construct validity. What follows contains, firstly, a review of the descriptive statistics including the cognitive ability measures of the test battery used in the participants' assessments. This is followed by the results of the simulations. The simulations mimic real world executive selection situations in which the selection criteria are somewhat 'fuzzy' in nature (Einhorn & Hogarth, 1975), as described in Chapter 6.

8.3.1 Descriptive Statistics

The psychometric measures used in this part of the analysis included cognitive ability measures as well as the personality measure. Table 20 shows the bivariate correlations between the measures used in this part of the research programme. In evaluating correlational effect sizes Cohen (1992) uses the following rules of thumb for the effect size impact of correlation coefficients - .1 (small), .3 (medium), .5 (large).

Table 20
Correlations between the Predictor measures used in the assessments

	AH4	Raven	N	E	O	A	C	BIDR-IM
AH4	-							
Raven	.69**	-						
N	-.09*	-.14**	-					
E	-.001	-.03	-.35**	-				
O	.17**	.12*	-.03	.31**	-			
A	-.03	-.03	-.30**	.14**	.15**	-		
C	.08	.03	-.60**	.34**	-.04	.28**	-	
IM	-.07	-.03	-.38**	.06	-.094	.35**	.40**	-

Notes. *Correlation is significant at the 0.05 level (2-tailed). **Correlation is significant at the 0.01 level (2-tailed). C – Conscientiousness, N – Neuroticism, E – Extraversion, O-Openness, A – Agreeableness, IM – bespoke BIDR-IM measure

Fifty-nine participants had a bespoke BIDR-IM score of 12 or greater. The correlations between the bespoke BIDR-IM scores and the personality dimensions are consistent with earlier research findings with bespoke BIDR-IM (Connelly & Chang, 2012), and three of the Big Five have medium intercorrelations that were statistically significant. The bespoke BIDR-IM scores for participants were tested for normality and while they were slightly positively skewed - coefficient of skewness .32 - the distribution of scores was normal within an acceptable level of skewness. It was positively skewed but approximately symmetrical, with skewness coefficient of .58.

8.3.2 Simulation Results

The outcome of the simulations for sets of 3 and 5 job candidate finalists, i.e. those who are on a selection short list, are shown in Table 21. The simulations were run on the basis that the finalist with the highest predictor score, or composite predictor score, will always be selected as the ‘winner’ from among the set of finalists. The proportion of simulations containing fakers (defined as those with a bespoke BIDR-IM score 12 or higher) refers to the number of simulated short lists which contain at least one faker out of the total 2000 simulations. The second last column shows the number of occasions when a faker is selected as the winner from the set of finalists as a proportion of the total number of simulations.

The results are in line with what would be expected in that the proportion should reduce as the selection ratio – defined as the number of finalists from whom a ‘winner’ is selected - reduces. There is little difference between the proportions of simulations containing fakers in all cases for a given selection ratio, regardless of the selection criterion or criteria used on which to base the finalist ranking. While fakers always have some likelihood of being selected the proportion of fakers selected was low in all cases, varying between 1 in 9 and 1 in 11 of the simulations. Fakers were around 20% more likely to be selected with the selection criteria do not include the cognitive ability measure, the AH4. Selection based on the AH4 only was used as the baseline criterion for selection against which each of the other criterion for selection were evaluated. This can be seen in the last column of Table 21.

Table 21

Simulation results of proportion of simulations containing Fakers, and the proportion of times a Faker is selected

Number of Finalists in Set	Predictors Used for Selection Criteria	Proportion of Simulations containing Faker	Absolute Difference from lowest	Proportion of Times Faker Selected	Absolute Difference from AH4 Simulation
3	AH4 Only	.660	.018	.105	-
3	C Only	.664	.022	.100	.005
3	C/N/AH4	.642	-	.111	.006
3	C/N/2.5AH4	.654	.012	.109	.004
3	All No A	.663	.021	.107	.002
3	C/N/E	.662	.020	.131	.026
5	AH4 Only	.494	.011	.090	-
5	C Only	.481	.002	.088	.002
5	C/N/AH4	.511	.028	.091	.001
5	C/N/2.5AH4	.490	.007	.105	.015
5	All No A	.498	.015	.102	.012
5	C/N/E	.483	-	.115	.025

Notes: AH4 – the cognitive ability measure, C – Conscientiousness, N – Neuroticism, E – Extraversion, A – Agreeableness. ‘All No A’ includes the scores for AH4, Raven’s, N, E, O, and C.

The next table shows that reducing the selection ratio, as defined above, would reduce the proportion of fakers. The comparisons are for four different selection ratios using the same selection criteria. The interesting finding is that the impact of reducing the selection ratio does not greatly change the incidence of ‘false positives’, which is defined as the selection of a faker from the set of finalists.

Table 22

Effect of number of Job Applicants in Selection Set on proportion of simulations containing Fakers, and the proportion of times a Faker is selected

Number of Finalists in Set	Proportion of Simulations containing Faker	Proportion of Times Faker Selected
3	.642	.111
4	.587	.105
5	.511	.091

To examine the extent to which the matched pairs rank order of the participants in the field study varied the next table looks at the rank order correlations for the set of three finalist case (Table 23). The 443 participants' rank orders were compared using their standardised scores on whatever single criterion or composite criterion was used in the simulations. The results showed that when a composite of three Big Five personality dimensions are used and then compared with selection based on the cognitive ability measure, AH4, there was no correlation between the finalists selected as 'winner' from the simulated sets of finalists. Selection based on solely on personality resulted in very different selection decisions to selection based solely on cognitive ability.

Table 23
Effect of Predictor Set on Rank Order of Job Applicants

Predictors Used for Selection Criteria	Spearman Rank Order Correlation with AH4 only Simulation
AH4 Only	--
C/N/AH4	.75
C/N/2.5AH4*	.94
All No A	.87
C/N/E	-.02

Notes: AH4 – the cognitive ability measure, C – Conscientiousness, N – Neuroticism, E – Extraversion, A – Agreeableness. ‘All No A’ includes the scores for AH4, Raven’s. N, E, O, and C.*AH4 weighting in composite increased from 1 to 2.5.

Finally, the impact of using minimum cut-off scores as hurdles in a non-compensatory, multiple hurdle, model for the selection criteria was examined (Table 24). The hurdles used were the mean score for the sample on each of the predictors used for selection criteria. As was to be expected the proportion of Fakers selected increased. This was also true when using a bespoke BIDR-IM score of 12 as the basis for dichotomization and using the bespoke BIDR-IM score as the selection hurdle.

Table 24

Effect of Different Cut-off Hurdles on proportion of simulations containing Fakers, and the proportion of times a Faker is selected

Number of Finalists in Set	Basis of Selection Criteria for Predictor Cut-off Hurdle	Proportion of Simulations containing Faker	Proportion of Times Faker Selected
3	All applicants with scores above Cut-off scores on each predictor	.619	.138
3	Fakers defined as those with BIDR-IM score of 12 or higher	.520	.146

Even though the proportion of participants with a bespoke BIDR-IM score of 12 or higher was 10.8% of the total number of participants, the proportion of Fakers selected was higher than this regardless of the cut-off hurdle or hurdles used as the basis for selection.

8.3.3 Rosse, Stecher, Miller, and Levin Study comparison

To further address the extent to which faking good was captured in the field study, the next table contains comparisons between the bespoke BIDR-IM scores of participants and those published in the Rosse, Stecher, Miller, and Levin (1998), as well as the norm group contained in the manual for the BIDR.

Table 25

Means and Standard Deviations of the bespoke BIDR-IM scale for Different Groups

	All Participants in the field study	Participants with BIDR- IM < 12	Participants with BIDR- IM = or > 12	BIDR-IM Rosse et al. (1998) Incumbents	BIDR-IM Rosse et al. (1998) Applicants	BIDR Norm Group
Mean	7.34	6.30	14.27	7.5	11.4	6.7
Std. Dev	3.79	2.75	1.95	3.0	4.1	4.0
Number	443	396	47	73	197	441

As previously mentioned in Chapter 7, Rosse et al. (1998) used both the NEO PIR and the bespoke BIDR-IM scale in their study. By using data from the Rosse et al. (1998) paper it was possible to make between participant comparisons using job applicants from this study and job incumbents from the Rosse et al. (1998) study. Comparing applicants and incumbents is one of the accepted research techniques used in evaluating faking good. Without such a comparison the question of the validity of the faking good assessment of the participant from the Managerial field study sample used in this research can be questioned. The comparison with the BIDR norm group sheds some light on how representative of the general population the participant group used in the research programme was. This issue was referred to in Chapter 6 in connection with restriction of range issues. The comparative descriptive statistics from the three sources are shown in Table 25.

Table 26 contains effect size statistics for a range of two group comparisons using Cohen's 'd' measure of effect size. In the table the 95% confidence intervals for

effect sizes which contain zero are shown in bold italics, and those that do not contain zero are underlined. The sign of the effect size is included in order to show the direction of the effect, if any. The effect size comparison between participants and the Rosse et al. (1998) Incumbents suggests that the faking good warning achieved the desired research strategy effect in that the participants and Rosse Incumbents are similar. There was a large effect size difference between the participants and the Rosse Applicants. The effect size comparison with the norm group was also low but the confidence interval did not include zero. All of the other comparisons had very large effect sizes except for the comparison between participants scoring less than 12 and 'Rosse Incumbents'.

Table 26

Effect sizes for Mean Score Differences in Impression Management scores between groups

Impression Management Score Comparison between groups	Effect Size 'd'
All Participants and Rosse Incumbents	<i>-0.04</i>
All Participants and BIDR-IM Norm Group	<u>0.16</u>
All Participants and Rosse Applicants	<u>-1.29</u>
Participants with BIDR-IM 12 or higher, and Rosse Incumbents	<u>2.63</u>
Participants with BIDR-IM 11 or less, and Rosse Applicants	<u>-1.56</u>
Participants with BIDR-IM 11 or less, and Rosse Incumbents	<u>-0.43</u>
Participants with BIDR-IM 12 or higher, and Rosse Applicants	<u>1.82</u>

Notes: Cohen's 'd' in italics – the 95% Confidence Interval includes zero. 'd' underlined – the 95% Confidence Interval does not include zero.

The results provide support for the dichotomisation carried out for the purposes of analysis in that those participants with a bespoke BIDR-IM score of 12 or higher were very different to the Rosse Incumbents and Rosse Applicants in their bespoke BIDR-IM score.

To briefly summarise the contents of the different results sections of this chapter – the various analyses of the results showed that there is good technical inferential evidence that the results support the central hypothesis of this research programme. Therefore the hypothesis, H1, that there are two uncorrelated higher order level factors, superordinate to the Big Five, which are not methodological artefacts was supported in the sample of participants with a bespoke BIDR IM score less than 12. Support for H1 means that hypothesis H2 - there is a single higher-order factor, the General Factor of Personality (GFP) superordinate to the Big Five – was rejected. Consequently, this finding was taken to support the hypothesis that an assessment procedure that includes a formal verbal warning was effective in, at least, minimising faking good. In addition, the bespoke BIDR-IM measure was shown to have played a useful and meaningful role in helping to dichotomise fakers from non fakers. This was of value in evaluating the consequential selection unfairness effect of some fakers not being eliminated from the selection process. Finally, the inferences from using the personality measure as well as the procedure used were shown to be invariant in the two field study samples. These findings will be explored in depth, largely from a theoretical perspective, in the next chapter.

Chapter 9

Discussion

Personality measures, such as the NEO PI-R omnibus personality measure, are widely used in applied settings for employee selection purposes. Arising from this there is the substantive and important theoretical issue of whether or not such measures lack strong (Kane, 2001) construct validity because of socially desirable responding due to faking good. There is also a fundamental question of major importance for practitioners to be answered. The question is whether these measures can be relied upon to arrive at optimal selection decisions in such applied contexts (Morgeson, Campion, Dipboye, Hollenbeck, Murphy, & Schmitt, 2007). The answers to these two critical questions, from both an applied and theoretical perspective, have significant implications for the ability of organisations to fully implement high performance work practices (Huselid, 1995).

The primary objective of this research programme was therefore to answer the two questions of the construct validity of personality measures and the appropriateness of their use in personnel selection by investigating the construct validity of the NEO PI-R in applied settings. In this research endeavour, the theory underlying the concept of construct validity (Loevinger, 1957; Messick, 1995) was relied upon. The achievement of the primary objective rested upon addressing the final three of the five issues listed in Chapter 1 in the actual field studies of the Managerial and Validation samples used in the research programme. The first two issues were the subject of the theoretical background reviewed in Chapters 2, 3, and 4. The remaining three issues were the subject of the actual field studies research:

1. Does the NEO PI-R yield construct validity in applied settings such as personnel selection?
2. Does faking good occur in such applied settings and if yes, can a formal warning reduce or even eliminate faking good?
3. Can the bespoke impression management measure that was used detect those who fake good despite the warning?

The empirical findings of the research programme relate to these three issues. The programme successfully achieved its primary objective by initially seeking to establish the internal structural aspect of the construct validity (Loevinger, 1957; Messick, 1995) of the NEO PI-R personality measure in high stakes employee selection contexts. It did so by evaluating the construct validity of this particular measure when used together with a formal verbal warning about the detection of lying by job applicants in the form of faking good when responding to item stems in the NEO PI-R. This was of vital importance to the research objectives and questions because of Messick's (1995) admonition that "the structural aspect appraises the fidelity of the scoring structure to the structure of the construct domain at issue" (p. 745). This required the determination of the factor structure of the NEO PI-R in the participants' samples and then comparing this structure with extant, construct valid, research.

The primary reason for establishing the factor structure, as part of the construct validation process, is that it made it possible to subsequently carry out the overall evaluative judgment (Messick, 1995) of the accuracy of the primary inference made in the research programme. The inference from this initial step was that the use

of a formal warning about the detection of faking good was effective in at least minimising faking good on the NEO PI-R personality measure in high stakes employee selection contexts. This inference was made possible by a comparison of the results of the investigation of the NEO PI-R higher order factor structure in this research programme with results from extant multitrait-multimethod (MTMM) studies. These studies were designed to account for common method variance (CMV; see Campell & Fiske, 1959). Specifically, they have shown that the Big Five higher order factors of Stability and Plasticity are either not correlated or only slightly correlated (Anusic, Schimmack, Pinkus, & Lockwood, 2009; Biesanz, & West, 2004; Chang, Connelly, & Geeza, 2012; DeYoung, 2006; Gnambs, 2015).

In summary, the results of the present research programme first and foremost showed that the primary objective of the research programme was achieved. The objective was to answer the question – can an assessment procedure be devised which allows construct valid inferences about candidates to be made when the NEO PI-R is used in high stakes employee selection contexts? This substantive research question posed was answered in the affirmative i.e. a formal warning does, at least, minimise faking good. The results of the research programme also showed that a sustainable defence can be made for the use of the bespoke BIDR-IM impression management measure for detecting the occurrence of faking good among job applicants in high stakes assessment contexts. Both of these findings have important theoretical and applied implications which will be discussed in this chapter. These include the question of what is the higher order structure of the Big Five dimensions of personality, and is there a need to use assessment procedures that deal with the possibility of job applicants faking good.

The major contribution of the findings arising from this research programme is that it showed that the use of a formal warning about the detection of impression management in the form of faking good by job applicants when completing a self-report personality measure was effective in eliminating faking good in some applicants. The formal warning arguably eliminated the occurrence of faking good by those who scored below a cut-off score on the impression management measure used. It is argued that the results of the research programme showed that the procedure followed in the field studies achieved this because the results obtained were consistent with those extant MTMM studies of the higher order structure of personality. This finding confirmed the hypothesis that the formal warning was effective in at least minimising CMV due to faking good. This is the first study to show that formal warnings are effective with applicants in high stakes employee selection situations when using the NEO PI-R.

In addition McCann, Ziegler, and Roberts (2012) state that “social desirability scales ... are inappropriate both for applied purposes and for the purposes of researching the mechanisms and processes underlying faking behavior” (p. 315). It is argued that the research findings of this programme also showed that a case can be made for the construct validity of the bespoke version of the BIDR-IM measure (Paulhus, 1984) used in the research programme to detect faking good by job applicants.

The present chapter is structured as follows: first, research findings pertaining to the three critical issues enumerated at the start of this chapter to be addressed in this thesis are reviewed. Section 9.1 focuses on the adequate establishment of construct validity of psychological instruments and the impact of faking good on the same, thereby addressing questions 1, and 2 respectively. Section 9.2 answers question 3,

namely, the use of the bespoke BIDR-IM measure to deal with the consequential aspect of strong construct validity (Messick, 1995). Arising from the research findings Section 9.3 addresses the theoretical aspects of the underlying psychological determinants of faking good, and its prevention. The limitations of the research, and some suggestions as to future research, are then examined in Section 9.4. Finally, the conclusions arrived at are discussed in Section 9.5.

9.1 Establishing the Construct Validity of the NEO PI-R

The first issue in the list of research questions deals with the construct validity of the NEO PI-R when used in high stakes employee selection situations. The establishment of construct validity, as theoretically defined in a series of landmark papers (Cronbach & Meehl, 1955; Loevinger, 1957; Embretson, 1983; Messick, 1995), is essential before any credence or valid inference can be attached to the scores obtained on a personality measure, such as that used in this research programme. Additionally and just as important, substantive trait variance can be contaminated by CMV in self-report measures and this must be accounted for before making inferences about scores on self-report measures (Campbell & Fiske, 1959).

The results of Section 8.2 of the previous chapter show that the confirmatory factor analysis (CFA) findings of the field studies demonstrate support for the existence of a two uncorrelated higher order factor model of the Big Five, as measured by the NEO PI-R, consisting of Stability and Plasticity as proposed by Digman (1997). As already pointed out the answer to this question was essential to

the process of establishing whether or not the formal warning at least minimised faking good. The research findings also make an important contribution to the debate concerning the higher order structure of the Big Five dimensions of personality.

9.1.1 The Higher Order Structure of the Big Five

The conclusion that the two higher order factors of the Big Five as measured by the NEO PI-R in the research programme are not correlated is important. In addition, no substantive evidence was found for the existence of the General Factor of Personality (GFP) advocated in Rushton and Irwing's (2008) model of the higher order structure of the Big Five. It is argued that the analyses carried out showed that the two latent higher order factors of Stability and Plasticity (DeYoung, 2006) were found to be uncorrelated after controlling procedurally for CMV in the form of faking good in the field study samples. This finding allows the inference to be made that the formal warning at least minimised the detrimental effect of CMV in the personality assessment of participants in the field studies. This is because the field monomethod studies of the research programme replicated the findings of extant MTMM studies (Anusic et al., 2009; Biesanz & West, 2004; Chang et al., 2012; DeYoung, 2006; Gnambs, 2013). These MTMM studies found that Stability and Plasticity were either uncorrelated or the correlation, if any, was very low when CMV was taken into account in the MTMM evaluation. The monomethod study of the two participant samples of this research programme found, by inference, that Stability and Plasticity were not correlated. It is the first monomethod study to do so arising from the fact that socially desirable responding was controlled for by the procedural use of the formal warning. This is an important finding.

The use of MTMM studies was shown by Campbell and Fiske (1959) to be an invaluable methodological technique for establishing the construct validity of self-report measures. It does so by largely separating method variance from substantive variance due to the latent traits in self-report measures, such as the NEO PI-R. Hence a similar finding from this research programme was essential to establish the internal structural aspect of the construct validity of the inferences arising from the use of NEO PI-R in the context of the field studies. If Stability and Plasticity are not correlated, as was found in the present research, then it can be reasonably inferred that CMV arising from socially desirable responding (Chapter 4) in the form of faking good was at least minimised as a result of the formal warning used in the personality assessments of participants (Chang et al., 2012; DeYoung, 2006).

Support for the existence of a GFP, and/or the correlated higher order factors Stability and Plasticity, comes largely from monomethod studies the results of which may have been contaminated by artefacts such as CMV as Chang et al. (2012) and others (Davies et al., 2016; Gnambs, 2013) point out. For this reason, a number of MTMM studies have raised serious questions regarding the existence of a GFP (Anusic et al., 2009; Biesanz & West, 2004; Chang et al., 2012; DeYoung, 2006; Gnambs, 2013). The findings of the monomethod field studies of the Managerial sample and the Validation sample of this research programme are consistent with the findings of these MTMM studies in this regard. For example, similar to what was found in this research programme (Model 4 of Table 12 in Chapter 8), Chang et al. (2012, p. 419) also found that their two higher order factor meta-analytic CFA MTMM model – the gold standard MTMM methodology - was an improper model which did not yield an admissible solution because of a negative residual variance. This supports the finding of the field studies of this research programme. This

consistency of findings between the research programme and Chang et al.'s (2012) CFA findings is evidence that the conclusion reached that Stability and Plasticity are not correlated in the two field study samples of this research programme.

In their article Chang et al. (2012) expressed the opinion that in order to test the competing hypotheses of the hierarchical structure of personality a multitrait approach is necessary. This was also the view of Anusic et al. (2009) who felt that firm conclusions about the higher order structure of the Big Five could only be arrived at using an MTMM approach. However, Johnson, Rosen, and Djurdjevic (2011b), in their investigation of the construct validity of Core Self Evaluations, showed that a monomethod study of higher order constructs can be as effective in investigating construct validity of a multidimensional construct. However, CMV must be controlled for in the monomethod study, as recommended by a number of researchers (Podsakoff et al., 2012; Richardson Simmering, & Sturman, 2009; Spector, Rosen, Richardson, Williams, & Johnson, 2017; Williams, Hartman, & Cavazotte, 2010). Consistent with this research, the monomethod field studies of the present research programme also show that it is possible to test the internal structural component of construct validity of both the two-factor hierarchical model of the Big Five and the single GFP model by using a procedural approach designed to control for CMV. This, in turn, permits the use of a statistical procedure to evaluate the effectiveness of the procedural controls in minimising CMV. The statistical procedure used in this research included some novel elements. Firstly, by using the two aspects of Extraversion (DeYoung, Quilty, & Peterson, 2007) as indicators in a number of the CFA models tested the problem of underidentification was dealt with. Secondly, a novel element was introduced in the analysis when using a Chi squared test to compare the factor

loadings for a two higher order factor model, each factor having three indicators, in order to determine the source of CFA model misfit.

Interest in the concept of a GFP owes much to Musek (2007) whose evidence for a GFP was based on monomethod studies. Musek's research did not follow the recommendations for controlling for CMV by Podsakoff, MacKenzie, Lee, and Podsakoff (2003). According to Danay and Ziegler (2011), "With monomethod studies, that is, mono-rater studies, as forwarded by Musek (2007), for example, this problem has not been overcome (Anusic, Schimmack, Pinkus, & Lockwood, 2009; DeYoung, 2010) because it is impossible to disentangle variance due to trait and variance due to bias" (p. 561).

The contribution of this research programme to theory is twofold, and can be summarised as follows:

- 1) The empirical findings show that even in a monomethod study it is possible to control for the variance due to CMV by following the recommendations of Podsakoff et al. (2003, 2012). This finding of the field studies is consistent with nearly all of the extant MTMM studies that investigated the higher order structure of the Big Five. This finding is consistent with the theory underlying MTMM analysis (Campbell & Fiske, 1959; McDonald, 1999)
- 2) Although the determination of the higher order structure of the Big Five was not the primary objective of this research programme, the finding that Stability and Plasticity are not correlated, in participants with a bespoke BIDR IM score less than 12, is an important contribution to the on-going theoretical debate concerning the higher order structure of the Big Five.

While ideal from a methodological perspective, the findings from the research programme clearly show that it is not necessary to carry out an MTMM study to control for error provided the recommendations of Johnson et al. (2011b) and Podsakoff et al. (2012) are followed. This finding is particularly important from a practical perspective as it is typically not feasible to have multiple raters to assess the personality traits of external job applicants in selection settings (Morgeson et al., 2007). The inference from the research finding that it was unlikely that there was a significant correlation between Stability and Plasticity is important. It provides good evidence that methodological artefacts in the form of socially desirable responding can substantively be the cause of the *shared* variance underling the Big Five dimensions loading on Stability and Plasticity. Monomethod studies that have been published in support of the existence of a GFP did not address this issue (Chang et al. 2012; Comensoli & MacCann, 2013).

Evidence, in this research programme, for a lack of support for a GFP comes from the fact that a) the conclusion arising from the poor factor loading results of the CFA evaluation of Models 1 and 2 of Table 12 in Chapter 8, and b) that by differentiating between the two aspects of Extraversion it was possible to compare the two factor higher order model with the GFP model, even though neither CFA model was meaningful because of very poor fit indices. The results of the analysis of the ‘just identified’ model of Section 8.1.3.2 in the previous chapter provide further support for this conclusion. It was only by using the two ‘aspects’ of Extraversion (DeYoung, Quilty, & Peterson, 2007) as indicators of Plasticity in the analysis of the results that this latent factor had the necessary, for CFA identification purposes, three indicators rather than just two.

As already mentioned, these findings from the field studies of this research programme are consistent with extant research from MTMM studies, but not with extant monomethod studies. To deal with the same problem arising from two orthogonal higher order factors Gnambs (2013) points out that “As bifactorial models with five traits are ordinarily not identified” (p. 510) and he goes on to state that “the respective factor loadings were estimated using the Schmid and Leiman (1957) procedure” (p. 510). In testing a correlated higher order two factor model Gnambs in his MTMM study followed the Kenny, Kashy and Bolger (1998) heuristics by having equal factor loadings for the two indicators of Plasticity. Van der Linden, Bakker, and Serlie (2011) did find in a monomethod study, using a long form Big Five measure similar to the NEO PI-R, that the uncorrelated Stability and Plasticity models they tested did not provide acceptable CFA results due to either very poor fit indices or improper solutions due to Heywood cases. However, that particular study used a sample that combined selection and assessment participants and did not use a formal warning. The factor loadings for the GFP found were similar to those of Musek (2007) which did not separate substantive trait variance from CMV (Comensoli & MacCann, 2013).

Using equal loadings on the two indicators of Plasticity to deal with the identification problem did not yield a solution in this research programme. This model compared the loadings of two, three indicators, single factor models with the loadings of a combined model consisting of two factors each with three indicators (see Model 3 of Table 12, Chapter 8). It was only by dealing with the identification problem in the analysis of the results of this research programme by having three rather than two indicators loading on Plasticity that the problematic one (i.e. minimum of three indicators) of the three conditions that must hold for identification (Kline, 2011) was

satisfied. Furthermore, it is also worth noting that Brown (2006) points out that “even if the correlation between two latent factors is zero a solution could be obtained if the two latent factors were measured by three indicators each” (p. 70).

The results of CFA models tested are consistent with the findings of DeYoung’s (2006) MTMM study. DeYoung’s analysis did not yield a proper CFA solution without fixing the factor loadings of Openness and Extraversion to be equal. However, this approach did not yield a solution in the analyses of this research programme. DeYoung made the point that ideally his findings should be replicated using the NEO PI-R because, *inter alia*, the personality inventory assesses a wider range of domain content than the short form measures that he used. This present study extends DeYoung’s research by using the long form NEO PI-R as the personality measure in a single informant field study while procedurally controlling for bias due to CMV arising from socially desirable responding. It also sheds light on the effect of CMV on conclusions arrived at about the competing theories of higher order structure of the Big Five based on self-report personality measurement and supports the findings of the extant MTMM studies on the topic (Anusic et al., 2009; Chang et al., 2012; DeYoung, 2006; Gnambs, 2013). As stated before the support for the existence of a substantive trait GFP comes primarily from monomethod studies the results of which are easily distorted by CMV (Gnambs, 2013). An important contribution of the findings of this monomethod study research programme to the debate on the higher order structure of the Big Five is that its findings are consistent with the extant MTMM studies.

Critical to carrying out MTMM studies are peer ratings of participants by others. These ratings are necessary in order to carry out an MTMM study but are not easily or readily available in applied settings (Morgeson et al., 2007). The field

studies consisted of job applicants each of whom had received a formal faking good warning. They completed a psychometrically valid impression management measure, and both the long form omnibus NEO PI-R and NEO-PI3 have greater reliability compared to short form Big Five measures used in much of the extant research. These factors made it was possible to better examine the effect of socially desirable responding in the form of faking good on the higher order structure, if any, of the Big Five. In addition the setting for the research programme was an applied setting the context of which was the result of implementing a procedural remedy for putative CMV as recommended by a number of researchers (Johnson, Rosen, & Chang, 2011a; Johnson et al., 2011b; Podsakoff et al., 2003; Podsakoff et al., 2012). The participant samples used provided a monomethod setting in which the distorting effect of CMV arising from faking good could be controlled for. The use of the NEO PI-R, and its alternative the NEO-PI3, ensured that random measurement error due to unreliability of the personality measure was minimised compared to the short form Big Five measures used in many research studies (DeYoung, 2006).

In summary, the contribution of this research programme to the study of faking good in high stakes employee selection contexts is that 1) it is the only applied monomethod study of the higher order hierarchical structure of the Big Five which included both a pre-test faking good warning and the bespoke impression management measure. 2) It was also the only such study that used the omnibus NEO PI-R personality measure in combination with the warning and bespoke BIDR IM measure. Because the findings replicated the extant MTMM studies of the higher order structure of the Big Five (Anusic et al., 2009; Biesanz & West, 2004; Chang et al., 2012; DeYoung, 2006; Gnambs, 2013) it is argued, by inference, that the formal warning was effective in at least minimising faking good among job applicants. This

is of great practical significance. The dimensions of personality have been shown to predict factors such as job performance (Roberts, Kuncel, Shiner, Caspi, & Goldberg, 2007) and career progression (Judge, Higgins, Thoresen, & Barrick, 1999). Yet the full benefits of this body of research have not been achieved because of construct valid issues with self-report personality measures at the selection stage of recruitment (Griffith & Converse, 2012). The use of a formal warning has been shown to at least minimise this very important applied problem.

9.1.2 The Link between Faking Good and the Higher Order Structure of the Big Five

The question as to whether faking good occurs or not in high stakes employee selection situations is a fundamental theoretical one that must be addressed when evaluating the construct validity of the NEO PI-R. The discussion in the previous section of the findings of this research programme provided evidence that a detrimental effect on the construct validity of the NEO PI-R can occur in high stakes selection situations by not controlling for socially desirable responding in the form of faking good. In doing so the first issue raised in the list of research questions to be addressed in the three empirical questions listed at the start of this chapter was dealt with. By inference faking good was shown to be detrimental to the structural aspect of construct validity (Messick, 1995) of the NEO PI-R in high stakes employee selection contexts.

The question of whether these empirical research findings, which have been taken to support the hypothesis that the faking good warning was effective, were due

to the faking good warning or not are still somewhat outstanding This is because there is a putative alternative explanation in that attenuation due to measurement error arising from unreliability may have cancelled out the effect of error due to CMV (Brannick, Chan, Conway, Lance, & Spector, 2010; Lance, Dawson, Birkelbach, & Hoffman, 2010; Spector, 2006) However, in defence of the findings of this research programme Johnson et al. (2011b) also showed that the procedural remedies such as those used in this study, as suggested in Podsakoff et al. (2003) for minimizing CMV, are effective in minimising CMV. The balancing out effect of CMV by measurement error, explained in Lance et al. (2010), does not always occur. Measurement error was less of a problem in this field study, compared with many other studies on the topic, because the NEO PI-R and NEO-PI3 were used which are more reliable compared to a short form measure such as the NEO-FFI – average coefficient alpha of .89 compared with .77 (Costa & McCrae, 1992). In addition, the findings of the monomethod field study of this research programme replicated the results of most of the MTMM studies with the exception of that of Rushton and Irwin (2008). The latter found evidence for a GFP in their research but their findings have been questioned (Comensoli & MacCann, 2013). As previously reviewed, the MTMM studies separated substantive trait variance from method variance to arrive at an uncontaminated estimate of the correlation between Stability and Plasticity. The correlation between Stability and Plasticity found in the Chang et al. (2012) meta-analysis was a good, even if not perfect, estimate of the true correlation because of the corrections made for unreliability and restriction of range in their meta-analytic MTMM study. It is the comparison of the findings of the higher order structure of the Big Five of this research programme with the findings of extant MTMM studies that

provides the causal link between the faking good warning and the higher order structure of the Big Five.

The analyses of the field studies' results also showed that the failure to find an admissible solution in the five factor models with two higher order factors may not have been due to a failure to adhere to the Kenny, Kashy and Bolger (1998) heuristics for identification in non-standard CFA models i.e. those in which there are correlated errors (see Models 3 and 4 in Table 12, Chapter 8). This is because an admissible solution (with poor fit indices) was obtained for Model 6, a six indicator model with two higher order factors with no correlated errors included in the model tested. The analysis of Section 8.2 of the previous chapter confirmed DeYoung's (2006) finding of essentially uncorrelated higher order factors, and found no incontestable evidence to support the GFP findings of either Musek (2007) or Rushton and Irwing (2008). For the modified (i.e. six indicators) Big Five model tested in this research programme with a single factor GFP model the fit was very poor, as well as having some very poor factor loadings. The CFA results for Models 3, 4 and 6, when considered together, suggest that the reason the two factors with five indicators model was inadmissible was due to Plasticity and Stability not being correlated rather than a correlated error problem.

In their research DeYoung, Peterson, and Higgins (2002) found that Stability and Plasticity, extracted from the NEO PI-R in a monomethod study of university students, correlated at .45. This present study found, by inference, that the two latent factors are likely not correlated, which is consistent with DeYoung's (2006) later findings from his MTMM research using the short form BFI personality measure. It is arguable, based on the findings of this current study, that the DeYoung et al.'s (2002) and Chang et al.'s (2012) findings of low, rather than no, correlations between

Stability and Plasticity may have been due to not have fully accounted for all forms of method variance in their research (Le, Schmidt, & Putka, 2009; Spector, Rosen, Richardson, Williams, & Johnson, 2017). Some support for this view comes from the recent Davies et al. (2016) meta-analysis which showed that there is evidence for method variance due to idiosyncratic differences between the personality measures of studies included in the meta-analysis, rather than being due solely to CMV.

By relying on a research strategy using the same omnibus personality measure together with a faking good warning, we can infer from the results that this procedural strategy was able to eliminate, maybe completely, the effect of CMV due to socially desirable responding in the form of faking good. This is of great practical importance when assessing personality in the applied context of high stakes employee selection situations. This is because the participants in the field studies may not have engaged in deliberate impression management to an extent that would invalidate the vast majority of participants' responses to items in the omnibus self-report personality inventory used. Another possible explanation of the findings of the research programme is that the small correlation found in the DeYoung (2006) and Chang et al. (2012) studies is that given by the equation E2 on page 136 of Chapter 6 – CMV may not be uniform across the studies and, for instance, there may have been a correlation between the self-report method effect and a halo effect in the peer reports (Spector, 2017) in these two MTMM studies. This would not have been a factor in the field monomethod study of this research.

The CFA analysis of this study showed that, when CMV was procedurally controlled for, it was not possible to arrive at a unique solution when just two of the five Big Five dimensions loaded on one of the two uncorrelated higher order factors, as McDonald's (1999) general conditions for local independence in CFA predict.

This finding could explain why Hopwood, Wright, and Donnellan (2011) encountered several estimation challenges in their research and why a number of the modelling approaches that they analysed yielded inadmissible solutions. Despite their attempts to address Heywood cases and negative factor variances, the various CFA models that they analysed yielded a complicated pattern of results that failed to support the GFP hypothesis. This issue of a failure to satisfy the local independence condition in CFA using the dimensions of the Big Five as indicators loading on two higher order factors, when CMV is accounted, for raises doubts about the findings of a GFP in other studies that have examined the higher order structure of the Big Five without controlling for CMV. It was this CFA finding of the research programme that led to the conclusion that a formal warning was effective in eliminating faking good among those who scored below the cut-off score on the bespoke BIDR-IM measure.

In summary, the link established in this research programme between the use of a pre-assessment formal warning and the higher order structures of the Big Five was relied upon to establish the adequacy and appropriateness of the interpretation of the NEO PI-R test scores by participants in the two field studies. The interpretation arrived at was that the employment of a pre-assessment formal warning was effective in eliminating faking good for participants who scored below the cut-off score on the bespoke BIDR-IM measure. It is also very important to clarify what is the best method in applied situations for accurately estimating to the fullest extent possible a job applicant's standing on the Big Five dimensions. Thus the ability to fully account for faking good by job candidates was of importance, from a strong construct validity perspective. This requires an evaluation of the consequential aspect (Messick, 1995) of possible bias and unfairness in selection decisions. The achievement of this latter

objective relied upon using the bespoke BIDR-IM measure for the detection of faking good in the research programme.

9.2 Construct Validity of the Bespoke BIDR-IM Measure

The question of how to measure faking good is also critical to the broader research aim of establishing the construct validity of personality measures in high stakes situations. Specifically, the present research programme addressed this question by using a bespoke version of the IM scale of the BIDR-IM (Paulhus, 1984). The evaluation of its construct validity was based on a number of different strands of evidence. These were the Connelly and Chang (2016) meta-analysis of socially desirable responding measures, the effect of restriction of range effects, and the implications of Borghans, Duckworth, Heckman, and ter Weel's (2008) formula for the relationship between the manifestation of a trait and its latent and contextual determinants. It also included a cluster analysis based on Connelly and Chang's (2016) meta-analysis, as well as a comparison of the two dichotomised groups of participants in this study with applicants and incumbents of the Rosse, Stecher, Miller, and Levin (1998) research so as to evaluate the effectiveness of the dichotomisation between those designated Fakers and non Fakers in this research programme.

The results of the Cluster Analysis when coupled with the CFA findings of this research programme (Tables 8 and 9 in Chapter 8) suggest that faking good was minimised. Firstly, the inferred absence of a correlation between Plasticity and

Stability in the Managerial field study sample with a bespoke BIDR-IM score less than 12 provides support for the use of the bespoke BIDR-IM scale in detecting faking good in the research. This conclusion is based on the convergence with the findings of extant MTMM studies. In addition, participants in Cluster 1 of the Cluster Analysis results (Chapter 8) can be considered to include those most likely to fake good. This conclusion is based on the Connelly and Chang (2016) meta-analytic MTMM study findings which showed that socially desirable responding to measures of the Big Five was associated with scores on the BIDR-IM scale, as well as scores on measures of both Agreeableness and Conscientiousness.

The percentage of participants in Cluster 1 of the Cluster Analysis was 61.6% and arguably, based on the research reviewed earlier (Sections 4.1 and 4.2 of Chapter 4) this may represent an upper bound estimate to the number of participants that might have faked good to a greater or lesser extent in the absence of the formal warning. A case can be made for assuming that Cluster 2 contained those who are unlikely to fake, and that Cluster 1 contains those extreme fakers as well as those who might fake good to some extent. In their research Zickar, Gibby, and Robie (2004) found that there were three classes of respondents namely, honest, slight, and extreme faker clusters. The percentage of honest (Cluster 2) participants was 29.4% in this research programme which is too different to the average figure of 37.7% for honest participants in the Zickar et al. (2004) research. Zickar et al. used a different personality measure and a different clustering technique, namely, mixed model item response theory, in their research. For these reasons the Zickar et al. findings and the research findings of this research programme were taken to be supportive of each other.

The number of participants at each possible score contained participants from both clusters. Scores of 12 or higher on the bespoke BIDR-IM measure contained no participants from Cluster 2, the putative honest cluster. The effect sizes differences between the two dichotomised groups of participants – those with a score 12 or higher and those with a score less than 12 - were large. This provides further support for the viewpoint that the formal warning together with the bespoke BIDR-IM measure and the restriction of range effect all combined to minimise or eliminate faking good in all participants except those scoring 12 or higher. Based on Connelly and Chang's (2016) results, it is also arguable that a smaller effect size difference for Agreeableness and Conscientiousness would have been expected if the procedural use of the pre-assessment warning did not have an impact. These factors may have led to a relative minimising of the effect of Agreeableness and Conscientiousness, compared to the latent self-report method factor, on the bespoke BIDR-IM scores. The Borghans et al. (2008) functional formula on page 165 in Chapter 6 explains why this could happen. The manifestation of a trait depends on a number of factors which can vary in their impact depending on the context. In addition, the cluster analysis results also provide further support for the use of a cut-off score of 12 on the bespoke BIDR-IM measure as can be seen from the distribution curve in Figure 9 of Chapter 8, but this conclusion must be regarded as tentative due to the small sample of participants scoring 12 or higher on the bespoke BIDR-IM measure.

The NEO PI-R measure and the BIDR-IM scale were both used in the research carried out by Rosse, Stecher, Miller, and Levin (1998). A between participant effect size analysis for IM scores of participants in the present study and both the applicants and incumbents in the Rosse et al. (1998) study was carried out. The effect size results

(Table 24 in Chapter 8) showed that the applicant/participants in this study were very similar, with respect to scoring on the IM measures used, to incumbents in the Rosse et al. (1998) study. Those designated as Fakers in the present study scored much higher than the Rosse et al. (1998) incumbents – effect size of +2.63. In addition, there was a moderate effect size difference (-.43) between the Rosse et al. incumbents and those who were designated as non-Fakers by the dichotomising based on using a bespoke BIDR-IM cut off score of 12 or higher to designate participants as Fakers in the present study.

For these reasons, it is argued that it can be concluded from the empirical results and theoretical argument (Chapter 6) that the use of the bespoke BIDR-IM measure together with the formal warning in the field studies of this research programme did succeed in at least minimising the effect of faking good in the participant samples. As a consequence, defensible conclusions could be drawn from the results of the research programme. Executive high stakes selection is usually the outcome of selecting the successful candidate from a short list (Highhouse, 1998; Hollenbeck, 2009; Ones & Dilchert, 2009). The importance of decision making by executives in determining organisational outcomes (Huselid, 1995; Ones & Dilchert, 2009) gives added importance to the need for a construct valid measure of faking good in executive selection situations that is based on the six aspects of Messick (1995). The consequential aspect of construct validity arising from the results of the research programme is examined in some detail in the next sub-section, because of its role in possible bias and unfairness when selecting candidates from a short list.

9.2.1 The Practical Implication of Construct Validity

The Monte Carlo simulations of this research programme were designed to investigate the consequential effects of the occurrence of faking good, in keeping with Messick's (1995) delineation of the six aspects of construct validity. Specifically, the findings show that in spite of taking procedural precautions (Johnson, Rosen, Chang, Djurdjevic, & Taing 2012; Podsakoff, MacKenzie, & Podsakoff, 2012) to minimise faking good in the selection processes, this behaviour can still occur to a worrying degree among applicants, contrary to what some research has found (Ellingson, Sackett, & Connelly, 2007; Hogan, Barrett & Hogan, 2007). Faking good raises serious questions about the applied use of personality measures that do not include procedural controls for faking good in executive selection, and selection processes more generally, regardless of support for the use of personality measures from criterion related validity studies. This is a very serious applied issue. For example, in Ireland the equality in employment legislation, e.g. the Equality Act (2004), requires the use of objective measures in employee selection situations. The question, therefore, arises whether the NEO PI-R is an objective measure in high stakes employee assessment contexts. The occurrence of lying behaviour by candidates in the form of faking good must be dealt with from a strong construct validity perspective.

Making the wrong selection decision when it comes to executive selection can have serious consequences for an organisation because of the impact of executive decisions on an organisation. The consequences for an organisation of a bad selection decision are much greater in the case of senior executives than in the case of

employees lower down in the hierarchy. The CEO of an organisation determines strategy and strategy determines future outcomes (Amernic & Craig, 2010; Boddy, 2011; Lease, 2006; Padilla, Hogan, & Kaiser, 2007; Resick, Whitman, Weingarden, & Hiller, 2009; Singh, 2008; Stein, 2003). For example, narcissism in CEO's is positively related to strategic dynamism and grandiosity, and it engenders extreme and volatile organisational performance (Chatterjee & Hambrick, 2007). It follows from this that factors such as accuracy and fairness in testing has a role to play in detecting signs of sub-clinical narcissism and psychopathy - both of which have been shown to be related to the Big Five personality traits (Paulhus & Williams, 2002; Ruiz, Smith, & Rhodewalt, 2001) - among executives, as well as the more narrowly focused criterion related validity need to be considered. Low Agreeableness is common to both these syndromes as are high scores on some of the facets of Neuroticism (Jakobwitz & Egan, 2006; Ruiz et al., 2001). At the sub-clinical level, these syndromes can lead to poor outcomes in organisations (Boddy, 2005; Chatterjee, & Hambrick, 2007; Grijalva, Harms, Newman, Gaddis, & Fraley, 2013; Jonason, Slomski, & Partyka, 2012; Stevens, Deuling, & Armenakis, 2012). Self-report personality measures are widely used as part of the selection process by organisations (see Chapter 3). The weighting given to these measures in the final selection decision can vary among organisations but, nevertheless, they do play an important role in the selection decision process (Hollenbeck, 2009; Morgeson et al., 2007).

Where the selection decision consists of selecting one successful candidate from a small set of finalists the simulations carried out as part of the present research programme showed that the proportion of sets of finalists containing at least one Faker can be as high 2 in 3 in the case of sets of 3 finalists, or 1 in 2 when there are 5

finalists. This should be a very worrying finding for organisations and practitioners because of the odds of selecting a Faker in the absence of objective criteria for making the final selection decision. If the selection decision making process were potentially biased i.e. does not rely on pre-determined criteria rather than a selection procedure such as an unstructured interview, the odds of selecting a Faker should be a concern. This is because it is both unfair to the non-Fakers and yields negative outcomes for organisations in that unsuitable candidates may be selected. This finding is even more alarming when it is noted that procedural safeguards to minimise the incidence of faking good were taken in the field study and the Monte Carlo simulations took this into account. The simulations showed that regardless of the criterion used as the basis for the executive selection decision the occurrence of false positives, defined as selecting a Faker from the set of finalists, varied approximately between 1 in 9 and 1 in 11 for the three and five finalist sets respectively. This finding should be of concern to both organisations and practitioners alike. Arthur, Glaze, Villado, and Taylor (2010) found that the extent of faking good was over 30% for four of the Big Five dimensions of personality among job applicants without procedural controls to minimise the level of occurrence of faking good. The incidence of those deemed to be Fakers in this research programme was 10.8% in the Managerial field study. It is also noteworthy that the simulations showed that using cut-off hurdles in the selection criterion increased the odds of selecting a Faker to 1 in 7.

Another interesting finding from the simulations is that including ability measures along with personality measures, as the basis for the selection decision, the combination of measures used made very little difference to the proportion of Fakers selected even though the ranking of participants in the sample did change. This is contrary to what Peterson, Griffith and Converse (2009) found. Unlike Peterson et al.

(2009) in the simulations of this research programme, there was no difference in the proportions of Fakers selected whether the selection decision was based solely on either cognitive ability or Conscientiousness. They are not the same Fakers because of the almost zero correlation between cognitive ability and Conscientiousness. The present study's research also showed that, unlike Vasilopoulos, Cucina, and McElreath (2005), the use of a faking good warning did not result in either of the cognitive ability measures used in the battery of tests correlating with Conscientiousness. This is an important finding because it could be taken to show that those with higher ability scores were not engaging in deep semantic processing (Hauenstein, Bradley, O'Shea, Shah, & Magill, 2017) of the items in the NEO PI-R. However, this would require further investigation.

In summary, this review of the simulation results further showed that the occurrence of false positives, defined as Fakers selected, in executive selection is a persistent one regardless of the remedy used. The contribution of this part of the research programme on the construct validity of the NEO PI-R is that it highlights the bias in high stakes assessments in favour of Fakers. This is a major issue at an applied level which needs to be addressed. This is true whether it is a procedural one as recommended by Podsakoff et al. (2012) such as using a faking good warning, and/or using both cognitive ability and personality measures as the basis of selecting a 'winner' from a set of finalists. Arising from the results of this research one solution to the problem is to always use a formal faking good warning and, in addition, to use an impression management measure similar to the bespoke BIDR-IM measure used in this research programme to eliminate Fakers from the set of finalists. There is a major practical benefit arising from this approach to personality assessments in high stakes

employee selection contexts if Huselid's (1995) high performance work practices are to be fully implemented in organisations.

9.3 The Psychology of Faking Good

Faking good is a consequence of a number of factors (Ellingson, 2012; Griffith & Converse, 2012) that produce systematic differences in test scores that are not solely due to the traits themselves, thereby invalidating construct validity of NEO PI-R. There is a strong argument to be made that the psychological explanation for the low level of faking found in the Managerial sample of this research programme is to be found in the issue of moral hypocrisy (Section 4.2 of Chapter 4). The most recent book published on the topic of faking good does not contain any reference to the body of research on the topic of moral hypocrisy (Ziegler, MacCann, & Roberts, 2012). Yet the use of the pre-assessment faking good warning showed that the findings of Batson, Thompson, Seuferling, Whitney, and Strongman's (1999) experimental research was replicated in the two field studies, and was effective in reducing the incidence of faking good.

The theory and empirical evidence underlying the phenomena of Duval and Wicklund's (1973) objective self awareness and the illusion of transparency (Gilovich, Savitsky, & Medvec, 1998) also help to explain why a formal verbal warning should be expected to, at least, minimise the occurrence of faking good in high stakes selection situations. This is because of the requirement for introspection and self-evaluation on the part of the individual being assessed in self-report personality assessments (Holden & Book, 2012). The illusion of transparency refers

to a feeling that one's internal states are more apparent to others than is actually the case (Gilovich et al., 1998). Objective self-awareness (OSA) involves introspection and self-evaluation of oneself as an object (Duval & Wicklund, 1972). In a state of OSA individuals' attention is focused on themselves rather than how they appear relative to others (Morin, 2011) and this state has been found to successfully reduce moral hypocrisy (Batson, 2008). OSA can lead to individuals focusing on a perceived, self evaluated, discrepancy between the actual self and the ideal or ought self (Pryor, Gibbons, Wicklund, Fazio, & Hood, 1977). In high stakes personality assessments research has shown that some individuals attempt to reduce this discrepancy by faking good (McFarland, Ryan, & Ellis, 2002).

Participants in the field studies, when completing the NEO PI-R self-report personality measure, would have been focussed on the self (Hauenstein et al., 2017; Robie, Brown, & Beaty, 2007). As Hamilton and Shuminsky (1990) point out that "Each of the various methods of inducing self-focused attention, particularly Fenigstein and Levine's (1984), is similar to the process of completing a personality test. In all instances, subjects are asked to engage in activities that make them think about themselves and see themselves as an object" (p. 1301). This is very relevant to the research findings of this programme. The focus by participants in the field studies on, and real time awareness of, possible discrepancies between the actual and ideal self in the high stakes selection situation (Hauenstein et al., 2017) were arguably conducive to a state of moral hypocrisy, as a result of OSA, among the participants. There are two additional psychological factors, to consider when evaluating the role of the context of the test administration in the Managerial field study. These are, firstly, the role of obedience/compliance (Milgram, 1963) in determining whether participants did or did not fake good following the warning, even though people do

differ in the extent to which they are prone to blind obedience (Ent & Baumeister, 2014). The setting in which the Managerial sample of participants completed the battery of tests consisted of a one-on-one interaction between the test administrator and the participant. All of the participants were tested by the same administrator in the same setting in the same location. Secondly, the effect of the power imbalance between the administrator and each participant on compliance by the participant with the warning. Ent and Baumeister (2014) point out that the landmark Milgram obedience experiments have a positive aspect in that they demonstrated that individuals are prepared to overcome their personal proclivities in a situation where obedience to an authority figure is required. A total of 65% of participants in the original Milgram studies obeyed the administrator and went all the way in ‘shocking’ participants (Blass, 1999; Burger 2009). The corresponding figure for the Meeus and Raaijmakers (1986) study of administrative obedience was 91.7%. It is arguable that many participants in the field studies of this research programme were faced with an analogous dichotomous choice similar to that of the participants in the obedience studies – to fake good and increase their chances of being selected or to respond honestly and, possibly, lessen their chances. It is not unreasonable to argue that the warning given to participants brought about compliance with respect to not faking good in a high proportion of the participants. Obeying an authority figure is still a fairly strong social norm (Twenge, 2009).

In addition, there was also a power asymmetry in the Managerial field study between the administrator and the participants. The administrator has legitimate and expert power and, arguably, reward and coercive power (Keltner, Gruenfeld & Anderson, 2003), unlike the participants. A recent study by Hiemer and Abele (2012) showed that individuals in a power-less position, such as the participants in this field

study, were less likely to engage in risky behaviour compared with those who had power, such as the administrator. This demonstrated reluctance for the power-less to avoid risky behaviour, and the activation of inhibition-related tendencies (Keltner et al., 2003), also supports the view that in a high stakes selection situation such as that of the field study risky faking good was minimised. Blass (1991) discusses the effect of personality and situational effects on behaviour in the Milgram obedience paradigm and points out that obedience behaviour can vary as a function of situational manipulations and differ among individuals within the same setting. Roberts and Caspi (2001) point out that to expect people to behave the same across situations is psychologically nonsensical. The situational effect of the warning, coupled with the power-less role of participants, may have resulted in a high level of compliance and honest responding in this field study compared to a high stakes assessment without a warning.

In the Validation field study sample participants were required to tick a box on the screen before completing the NEO-PI3 acknowledging that they had read the warning. Shu, Gino, and Bazerman (2011) found that in their research participants' self-reported performance on a problem solving task, when they had the opportunity to respond dishonestly, was greatly reduced when they were required to read an honour code before participating in the problem solving task, and was eliminated when they signed the honour code after reading it. The authors concluded that by increasing moral saliency through having participants read or sign an honour code significantly reduced unethical behaviour and prevented subsequent moral disengagement – “a simple intervention, such as merely reminding actors about established moral codes, could counteract the effect of a permissive situation” (p. 344). Having a job applicant sign a written warning form before completing a

personality inventory may be a very effective way of eliminating faking good (Ayal, Gino, Barkan, & Ariely, 2015; Shu, Mazar, Gino, Ariely, & Bazerman, 2012).

The formal warnings used in the field studies may also have increased the salience of cognitive dissonance in those participants intending to fake good. This would have resulted in a change in attitude towards faking good in order to reduce psychological discomfort due to cognitive dissonance (Elliot & Devine, 1994). Participants' who might have faked good were faced with a potential loss after hearing the warning. Prospect theory (Kahneman & Tversky, 1979) also suggests that these participants may have been less likely, post the warning, to engage in the now risky behaviour of faking good because of a belief in the potential gain from honest responding compared with a putative more certain loss because of the warning.

To conclude did the present research programme answer the critical questions 1) Is the NEO PI-R construct valid in applied settings such as personnel selection? 2) Does faking good occur in such applied settings and if yes, can a formal warning reduce or even eliminate faking good? 3) Can the bespoke impression management measure that was used detect those who fake good despite the warning? The answers are yes.

Firstly, the empirical evidence reviewed from social psychology, behavioural economics, and industrial/organisational psychology is supportive of the viewpoint that faking good does indeed occur in high stakes employee selection situations in which the NEO PI-R was used. Secondly, the conclusion from the CFA results of models tested in the monomethod field studies is consistent with the findings of MTMM studies concerning the structure of personality, thereby supporting the hypothesis that a formal verbal warning in a monomethod context is effective in at least minimising faking good. This is a major contribution from the research

programme in helping efforts to ensure the accuracy and appropriateness of personality assessments using the NEO PI-R in high stakes employee situations. Finally, the bespoke BIDR-IM scale was shown to be effective in detecting faking good even though the level of proof was not absolute. In addition, the Monte Carlo simulations show that faking good can easily result in biased and unfair employee selection decisions.

9.4 Limitations and Suggestions for Future Research

Although the present monomethod field study using the NEO-PI-R along with a warning in a high stakes selection scenario offers a number of important advantages, it is not without its limitations. With regard to the analysis of construct validity of the NEO PI-R for those with a bespoke BIDR-IM score of less than 12, there are a number of limitations which should be mentioned. First, strictly speaking the findings of the research programme only apply to the NEO PI-R and the alternative version the NEO-PI3, and not necessarily to any other omnibus or short form Big Five measure, because even though these various measures are convergent they are not tau equivalent or parallel measures (Hopwood, Wright, & Donnellan, 2011; McDonald, 1999). Second, it may also be the case that the common variance underlying each of the higher order factors of Stability and Plasticity in this study is attributable to artefacts not controlled for such as item secondary loadings, acquiescence bias or positive and negative evaluative bias in the items (Davies et al., 2016). A third limitation is that the sample of participants was not a random sample of the population at large, so the generalisability of the results can be questioned, particularly in samples whose participants are not being assessed in high stakes situations. However,

both the Box's M Test and the exploratory factor analysis (EFA) of Section 8.1 did show that the intercorrelations between the Big Five were similar to the NEO PI-R norming sample, and the expected pattern of EFA loadings was recovered from the field study sample. This finding supports McCrae and Terraccino's (2005) position that the Big Five are universal human characteristics.

From a methodological perspective, strictly speaking CFA cannot be used as a 'confirmatory' technique when modifications of the original model abandon the theory driven confirmatory logic of CFA. This makes some of the results somewhat data driven and inductive like EFA. McCrae, Zondermann, Costa, Bond, and Paunonen (1996) made the point that in actual analyses of personality data structures that are known to be reliable showed poor fit when evaluated by CFA techniques and in their opinion this points to serious problems with CFA itself when used to examine personality structure. This may have been particularly so with the CFA using the facet scores of the Validation sample, where the Mardia coefficient greatly exceeded the recommended value. According to Rodgers (2010), if we are evaluating theories, and working in confirmatory mode, different models are specified and the best model is the one that fits the data best in relation to its complexity. CFA, in this regard, can be viewed as a tool for exploring the structure of general psychological attributes as accounted for by different putative models when it comes to evaluating the hierarchical structure of omnibus personality measures. This study did show by inference that the 'best' model among the competitors of the hierarchical structure of the Big Five was the one in which there are two uncorrelated factors at a level superordinate to the modified six dimensions model. Correlated error problems can occur, as Johnson's (1994) research suggests, arising from the fact that the Big Five facets of the NEO PI-R have both primary and secondary factor loadings which can

lead to correlated errors among the indicators of the latent constructs. CFA, however, can still be used as a tool with which to compare putative models and thereby test theory (Rodgers, 2010).

Extant CFA studies of the higher order factor structure of the Big Five formed the basis for the hypothesis tested i.e. for the warning to be shown to be effective the two higher order factors of Plasticity and Stability would have to be essentially uncorrelated. Recent advances in software arising from the use of exploratory structural equation modelling (ESEM) mean that EFA can now be used in a confirmatory manner (Asparouhov & Muthén, 2009; Marsh, Lüdtke, Muthén, Asparouhov, Morin, Trautwein, & Nagengast 2009; Marsh, Morin, Parker, & Kaur, 2014). With CFA the indicators are restricted to loading on a single factor. Indicator crossloadings that might exist between factors are constrained to be zero in CFA. These crossloadings, on the other hand, are freely estimated in EFA. However, these crossloadings might be important because requiring them to be zero typically results in inflated CFA factor correlations (Marsh et al., 2014). In practice in many applications modification indices are used to improve model fit in CFA by allowing for crossloadings. Essentially this means that the analysis is no longer strictly confirmatory (Brown, 2006; Kline, 2011).

ESEM can overcome this problem because as a more general framework for factor analysis it incorporates CFA and EFA as special cases. With ESEM, in addition to CFA measurement model parts, the EFA measurement model parts with factor loading matrix rotations can also be used. This means that a set of a priori model alternatives, as well as the hypothesis that was tested in the research programme of this thesis, can be subjected to testing (Marsh et al., 2014). These ESEM model alternatives can include crossloadings between the Big Five dimensions which were

shown in Chapter 3 to occur due to the secondary loadings of items in the NEO PI-R. This allows for model comparisons to be made based on chi square difference tests thereby lessening the dependence on goodness-of-fit indices rules of thumb in evaluating the models tested.

The effectiveness of measures such as the bespoke BIDR-IM scale in detecting faking good by job applicants when completing personality inventories has been questioned by several researchers (Ellingson, Heggestad, & Makarius, 2012). This was an issue in the research programme. In response to this a number of comments can be made. First, Connelly and Chang (2016) found that the BIDR-IM scale was found to more effectively tap self-report method variance than other socially desirable responding measures, even though it is also contaminated with loadings on Conscientiousness and Agreeableness. According to Connelly and Chang (2016), the BIDR-IM scale is not ‘entirely incapable’ of assessing an impression management response styles in self-report measures. Second, as a form of impression management, faking good in personality assessment is an example of moral hypocrisy in that it is a deliberate attempt to tailor one’s test responses to the demands of a particular testing situation. Lonnqvist, Irlenbusch, and Walkowitz (2014) showed that the impression management of those who deliberately engage in it is generally not accompanied by self-deception. While it may not be a perfect measure the BIDR-IM does nevertheless capture method variance due to deliberate impression management (Chang et al., 2012). The bespoke version used in this research programme was designed to take account of this deliberate faking good in the high stakes selection context of the research programme. Third, the findings of a ceiling and floor effect in the behavioural economics studies and other research (Borghans et al., 2008; Fischbacher, & Heusi, 2013; Hauenstein et al., 2017) do support the concept and use of a floor (or

ceiling) effect when applying a cut-off score in the analyses for the bespoke BIDR-IM measure.

The Monte Carlo simulation study also had some limitations. First, they were based on a convenience sample of applicants for executive level positions. This is not a random sample. Therefore, the sample used to generate the sets of finalists is not necessarily representative of the population of interest at large. However, Ones and Dilchert (2009) showed that there is substantial variance in cognitive ability and personality trait levels among executives which leads to a relatively acceptable level of variance that can be meaningfully analysed. Second, the participants' IM scores were dichotomised for the purpose of the analyses even though there was a slight positive skew in the distribution. The effect size comparisons with the Rosse et al. (1998) participants were of such a size suggest that this limitation should not invalidate the results. However, it should be pointed out that the Rosse et al. (1998) applicants and incumbents that were used for between participants' comparisons by Rosse et al. (1998) were for non executive positions. Yet in spite of these limitations the findings of the research programme are robust because the Managerial field study sample a showed similar pattern of intercorrelations as the norm group for the NEO PI-R, and the correlations between the ability and the personality measures used were in line with findings from other research studies (Ones & Dilchert, 2009; Schmit & Hunter, 2004).

Future research on this topic should examine the effect of different types of warning. The two field studies used different forms of a formal warning. In the Managerial field study, a verbal warning was used. For the Validation study, a written warning was used. The invariance analysis results suggest that there was no difference due to the form of the warning. There has been little research on this topic (Pace and

Borman, 2006). The effect of participants actually signing a written formal warning at the top of the written warning regarding the possibility of detecting faking good before completing a self-report measure such as the NEO PI-R should be investigated in applied settings. This is an issue that needs to be examined in future research in order to fully resolve the question of the higher order structure of the Big Five. Another issue that warrants further investigation is the effect of different methods of stimulating OSA. The presence of a mirror facing participants when completing personality measures merits investigation based on Batson et al.'s (1997) research findings. Another method for triggering OSA in the assessment process that could be investigated would be the visual recording of the assessment session as another method for triggering OSA. These methods have been used in the investigation of OSA in other contexts (Silva & Duval, 2001).

For future research the experimental paradigms of Ellingson, Heggstad, and Makarius (2012) and Fischbacher and Föllmi-Heusi (2013) could also be combined in a new experiment. The proposed experiment would be a within participant study. In the experiment, participants would be asked to complete a battery of tests that includes the IPIP-NEO Big Five measure, the bespoke BIDR IM, the NEO PI-R, and the Fischbacher and Föllmi-Heusi (2013) dice experimental paradigm, described in Chapter 4 on p.97. Following Ellingson et al. (2012) the accuracy of the NEO PI-R would be estimated by computing the difference in standard scores on the NEO-FFI and the IPIP-NEO Big Five measure as a baseline personality measure. The instructions would be worded so as to engender a natural motivation in the participants to engage in intentional distortion so as to create a favourable impression. The dice experiment of Fischbacher and Föllmi-Heusi, described in Section 4.2.3, is included in order to have a moral hypocrisy measure included in the experiment.

After the initial phase of the study participants would be retested using the NEO PI-R and bespoke BIDR IM under one of two instructional conditions. They would either be told that they were flagged because their initial scores were invalid, based on their bespoke BIDR IM score. Alternatively, they would be told that their data had been lost due to an administrative error. The outcome of interest would be whether retesting flagged individuals results in more accurate personality scores in the second assessment relative to the initial assessment. By comparing baseline scores with initial and retest scores on the NEO PI-R obtained in the motivating setting it should be possible to determine if retest scores are more accurate than initial scores. A comparison of those participants with scores above the cut-off on the bespoke BIDR IM measure with those participants with increases in the accuracy of their NEO PI-R scores could help to shed further light on the construct validity of the bespoke BIDE IM measure.

9.5 Conclusions

The extant research reviewed on the use of personality measures in the assessment of job applicants in high stakes contexts in the present thesis clearly indicates that faking good on personality measures is a real problem when it comes to the applied context of employee selection. Resolving this issue is of major practical importance in such a context. According to Drasgow, Stark, Chernyshenko, Nye, Hulin, and White (2012, p. 2), “intentional distortion can severely undermine the utility of measures for personnel selection”. By controlling for CMV in a high stakes employee selection setting, the higher order hierarchical structure of the Big Five

dimensions was shown in the research programme, by inference, to consist of two uncorrelated higher level factors – Stability and Plasticity – with Conscientiousness, Agreeableness and Neuroticism loading on Stability, and Openness and Extraversion loading on Plasticity. This finding was the methodological device that was used to evaluate the effectiveness of the assessment procedure followed for dealing with faking good in using the NEO PI-R in a high stakes employee selection context,

The procedural control of using a formal verbal warning to prevent or minimise lying in the form of faking by participants prior to completion the NEO PI-R omnibus personality measure in the field study was shown to be effective. This was because these two latent higher order factors, Plasticity and Stability, were found to be uncorrelated. This is of immense practical importance in the applied setting of high stakes employee selection situations. It, therefore, follows that a formal warning should always be used in such applied settings in order to ensure construct validity. The findings also provide support for Digman's theory (Digman, 1997) of two higher order factors superordinate to the Big Five, and found no substantive support for a General Factor of Personality at the apex of a hierarchical structure of personality. This is an important contribution to the debate on the higher order structure of the Big Five of personality (Anusic et al., 2013; Chang et al., 2012; Comensoli & MacCann, 2013; Gnambs, 2013). The present research is the first monomethod study to support the findings from extant MTMM studies that have investigated the higher order structure of the Big Five personality factors. This has a major practical advantage at the applied level because it can contribute in a substantial way to the achievement of Huselid's (1995) high performance work practices.

Using the bespoke BIDR-IM impression management measure the Monte Carlo simulations showed that basing the selection decision on subjective criteria e.g.

reliance on unstructured interviewing, the odds of selecting a Faker can be as high as 2 in 9 even when procedural controls to prevent faking are used. This is not a trivial finding. Even with criterion composites of multiple predictors the odds can be as high as 1 in 7. Relying on selection criterion composites that include cognitive ability and personality measures reduces the odds to around 1 in 10 for compensatory top-down selection models. This is still high because of the risks arising from selected an unsuitable candidate for an executive position. What is particularly worrying is that these results were found to be the case after procedurally controlling for the incidence of faking good. In the absence of such precautions the odds of selecting a Faker would have been much higher.

In conclusion, the present research, as well as the results of other scholars, suggests that construct validity is still an issue of concern in selection research (Ployhart, Schmitt, and Tippins (2017). As Landers, Sackett, and Tuzinski (2011) state that “Just assuming that faking is not a problem will not make it go away” (p. 210). Messick’s (1995) concept of construct validity, as well as that of other theorists, did not envision that the nomological net of the Big Five should be neglected simply because criterion validity studies are supportive of some of the inferences made about aspects of the construct validity of personality measures. Further investigation of these issues is advisable. This could be achieved by carrying out further research on the use of different forms of the formal warning and, also, the investigation of different methods that could be used to stimulate OSA.

REFERENCES

- Adair, C. (2014). Interventions for Addressing Faking on Personality Assessments for Employee Selection: A Meta-Analysis. *Thesis*. DePaul University
- Aluja, A., García, Ó., García, L. F., & Seisdedos, N. (2005). Invariance of the “NEO-PI-R” factor structure across exploratory and confirmatory factor analyses. *Personality and Individual Differences*, 38(8), 1879-1889.
- Amernic, J. H., & Craig, R. J. (2010). Accounting as a facilitator of extreme narcissism. *Journal of Business Ethics*, 96(1), 79-93.
- Anusic, I., Schimmack, U., Pinkus, R., & Lockwood, P. (2009). The nature and structure of correlations among Big Five ratings: The Halo-Alpha-Beta model. *Journal of Personality and Social Psychology*, 97, 1142–1156.
- Arbuckle, J. L. (2014). IBM SPSS Amos 21 user’s guide. Available: [f tp. public. dhe. ibm.com/software/analytics/spss/documentation/amos/21.0/en/Manuals/IBM_SPSS_Amos_Users_Guide.pdf](http://public.dhe.ibm.com/software/analytics/spss/documentation/amos/21.0/en/Manuals/IBM_SPSS_Amos_Users_Guide.pdf). Accessed.
- Arnau, R. C., Green, B. A., Rosen, D. H., Gleaves, D. H., & Melancon, J. G. (2003). Are Jungian preferences really categorical?: an empirical investigation using taxometric analysis. *Personality and Individual Differences*, 34(2), 233-251.
- Arthur, W., Glaze, R. M., Villado, A. J., & Taylor, J. E. (2010). The magnitude and extent of cheating and response distortion effects on unproctored internet based tests of Cognitive Ability and Personality. *International Journal of Selection and Assessment*, 18(1), 1-16.
- Ashton, M. C., Lee, K., Perugini, M., Szarota, P., de Vries, R. E., Blas, L. D., Boies, K., & De Raad, B. (2004). A six-factor structure of personality descriptive

- adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology*, 86, 356–366.
- Ashton, M. C., Lee, K., Goldberg, L. R., & de Vries, R. E. (2009). Higher order factors of personality: do they exist? *Personality and Social Psychology Review*, 13(2), 79-91.
- Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 16(3), 397-438.
- Ayal, S., Hochman, G., & Ariely, D. (2016). Editorial: Dishonest Behavior, from Theory to Practice. *Frontiers in Psychology*, 7.
- Ayal, S., Gino, F., Barkan, R., & Ariely, D. (2015). Three principles to REVISE people's unethical behavior. *Perspectives on Psychological Science*, 10(6), 738-741.
- Bacher, J., Wenzig, K., & Vogler, M. (2004). SPSS Two Step Cluster - a first evaluation. *Arbeits- und Diskussionspapiere* 2(2).
- Bäckström, M., Björklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality*, 43(3), 335-344.
- Bagozzi, R. P. (1993). Assessing construct validity in personality research: Applications to measures of self-esteem. *Journal of Research in Personality*, 27(1), 49-87.
- Bagozzi, R. P., & Edwards, J. R. (1998). A general approach for representing constructs in organizational research. *Organizational Research Methods*, 1(1), 45-87.
- Bandura, A. (2002). Selective moral disengagement in the exercise of moral agency. *Journal of Moral Education*, 31(2), 101-119.

- Bangerter, A., Roulin, N., & König, C. J. (2012). Personnel selection as a signaling game. *Journal of Applied Psychology, 97*(4), 719.
- Barger, S. D. (2002). The Marlowe-Crowne affair: Short forms, psychometric structure, and social desirability. *Journal of Personality Assessment, 79*(2), 286-305.
- Barkan, R., Ayal, S., Gino, F., & Ariely, D. (2012). The pot calling the kettle black: Distancing response to ethical dissonance. *Journal of Experimental Psychology: General, 141*(4), 757-773.
- Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*(1), 1-26.
- Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 81*(3), 261-272.
- Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: what do we know and where do we go next? *International Journal of Selection and Assessment, 9*(12), 9-30.
- Batson, C. D. (2008). Moral masquerades: Experimental exploration of the nature of moral motivation. *Phenomenology and the Cognitive Sciences, 7*(1), 51-66.
- Batson, C. D., Kobrynowicz, D., Dinnerstein, J. L., Kampf, H. C., & Wilson, A. D. (1997). In a very different voice: unmasking moral hypocrisy. *Journal of Personality and Social Psychology, 72*(6), 1335-1348.
- Batson, C. D., Thompson, E. R., Seufferling, G., Whitney, H., & Strongman, J. A. (1999). Moral hypocrisy: appearing moral to oneself without being so. *Journal of Personality and Social Psychology, 77*(3), 525-537.

- Bazerman, M., & Moore, D. A. (2012). Judgment in managerial decision making. Wiley.
- Bentler, P. M. (1995). *EQS structural equations program manual*. Multivariate Software.
- Berry, C. M., & Sackett, P. R. (2009). Faking in personnel selection: tradeoffs in performance versus fairness resulting from two cut - score strategies. *Personnel Psychology*, 62(4), 833-863.
- Bess, T. L., & Harvey, R. J. (2002). Bimodal score distributions and the Myers-Briggs Type Indicator: fact or artifact? *Journal of Personality Assessment*, 78(1), 176-186.
- Biderman, M. D., & Nguyen, N. T. (2009). Measuring faking propensity. *Paper in 24th Annual Conference of the Society for Industrial and Organizational Psychology*.
- Biderman, M. D., Nguyen, N. T., Cunningham, C. J., & Ghorbani, N. (2011). The ubiquity of CMV: The case of the Big Five. *Journal of Research in Personality*, 45(5), 417-429.
- Biesanz, J. C., & West, S. G. (2004). Towards Understanding Assessments of the Big Five: Multitrait-multimethod Analyses of Convergent and Discriminant Validity Across Measurement Occasion and Type of Observer. *Journal of Personality*, 72(4), 845-876.
- Biesanz, J. C., West, S. G., & Millevoi, A. (2007). What do you learn about someone over time? The relationship between length of acquaintance and consensus and self-other agreement in judgments of personality. *Journal of Personality and Social Psychology*, 92(1), 119-135.

- Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A Meta Analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, 14(4), 317-335.
- Blass, T. (1991). Understanding behavior in the Milgram obedience experiment: The role of personality, situations, and their interactions. *Journal of Personality and Social Psychology*, 60(3), 398-413.
- Blass, T. (1999). The Milgram Paradigm After 35 Years: Some Things We Now Know About Obedience to Authority. *Journal of Applied Social Psychology*, 29(5), 955-978.
- Block, J. (2010). The five-factor framing of personality and beyond: Some ruminations. *Psychological Inquiry*, 21(1), 2-25.
- Borckardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: a guide to case-based time-series analysis. *American Psychologist*, 63(2), 77-95.
- Boddy, C. R. (2005). The implications of corporate psychopaths for business and society: An initial examination and a call to arms. *Australasian Journal of Business and Behavioural Sciences*, 1(2), 30-40.
- Boddy, C. R. (2011). The corporate psychopaths theory of the global financial crisis. *Journal of Business Ethics*, 102(2), 255-259.
- Borghans, L., Duckworth, A. L., Heckman, J. J., & Ter Weel, B. (2008). The economics and psychology of personality traits. *Journal of Human Resources*, 43(4), 972-1059.
- Bornstein, R. F. (2011). Toward a process-focused model of test score validity: Improving psychological assessment in science and practice. *Psychological Assessment*, 23(2), 532-544.

- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111(4), 1061-1071.
- Brannick, M. T., Chan, D., Conway, J. M., Lance, C. E., & Spector, P. E. (2010). What is method variance and how can we cope with it? A panel discussion. *Organizational Research Methods*, 13(3), 407-420.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Press.
- Burger, J. M. (2009). Replicating Milgram: Would people still obey today?. *American Psychologist*, 64(1), 1-11.
- Burns, G. N., & Christiansen, N. D. (2011). Methods of measuring faking behaviour. *Human Performance*, 24(4), 358-372.
- Button, S. B., Mathieu, J. E., & Zajac, D. M. (1996). Goal orientation in organizational research: A conceptual and empirical foundation. *Organizational Behavior and Human Decision Processes*, 67(1), 26-48.
- Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Routledge.
- Byrne, Z. S., Peters, J. M., & Weston, J. W. (2016). The struggle with employee engagement: Measures and construct clarification using five samples. *Journal of Applied Psychology*, 101(9), 1201-1227.
- Caldwell-Andrews, A., Baer, R. A., & Berry, D. T. R. (2000). Effects of response sets on NEO-PI-R scores and their relations to external criteria. *Journal of Personality Assessment*, 74, 472-488.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105.

- Camerer, C. (1981). General conditions for the success of bootstrapping models. *Organizational Behaviour and Human Performance*, 27(3), 411-422.
- Caprara, G. V., Barbaranelli, C., & Borgogni, L. (1994). BFO — Big Five Observer. Manuale. Firenze: Organizzazioni Speciali.
- Caprara, G. V., Barbaranelli, C., & Zimbardo, P. G. (1997). Politicians' uniquely simple personalities. *Nature*, 385(6616), 493-493.
- Caprara, G. V., & Zimbardo, P. (2004). Personalizing politics. *American Psychologist*, 59(7), 581-594.
- 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.
- Carver, C. S., & Connor-Smith, J. (2010). Personality and coping. *Annual Review of Psychology*, 61, 679-704.
- Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology*, 56, 453-484.
- Cellar, D. F., Miller, M. L., Doverspike, D. D., & Klawnsky, J. D. (1996). Comparison of factor structures and criterion-related validity coefficients for two measures of personality based on the five factor model. *Journal of Applied Psychology*, 81(6), 694-704.
- Chambers, J. R., Epley, N., Savitsky, K., & Windschitl, P. D. (2008). Knowing too much using private knowledge to predict how one is viewed by others. *Psychological Science*, 19(6), 542-548.

- Chan, D. (2009). So why ask me? Are self report data really that bad. *Statistical and Methodological Myths and Urban Legends: Doctrine, verity and fable in the organizational and social sciences*, 309-336.
- Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology*, 82(1), 143.
- Chang, L., Connelly, B. S., & Geeza, A. A. (2012). Separating method factors and higher order traits of the Big Five: A meta-analytic multitrait-multimethod approach. *Journal of Personality and Social Psychology*, 102(2), 408 - 426.
- Chatterjee, A., & Hambrick, D. C. (2007). It's all about me: Narcissistic chief executive officers and their effects on company strategy and performance. *Administrative Science Quarterly*, 52(3), 351-386.
- Chiaburu, D. S., Oh, I. S., Berry, C. M., Li, N., & Gardner, R. G. (2011). The five-factor model of personality traits and organizational citizenship behaviors: a meta-analysis. *Journal of Applied Psychology*, 96(6), 1140-1166.
- Christiansen, N. D., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (1994). Correcting the 16PF for faking: effects on criterion - related validity and individual hiring decisions. *Personnel Psychology*, 47(4), 847-860.
- Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, 17(1), 31-43.
- Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement*, 70(5), 732-743.

- Clark, L., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309-319.
- Cohen, Jacob. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159.
- Cohen, Jacob. (1994). The earth is round ($p < .05$). *American Psychologist*, 49(12), 997-1003.
- Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *The American Journal of Medicine*, 119(2), 166.e7 -166.e16.
- Collins, J. M., Schmitt, F. L., Sanchez-Ku, M., Thomas, L., McDaniel, M. A., & Le, H. (2003). Can basic individual differences shed light on the construct meaning of assessment center evaluations? *International Journal of Selection and Assessment*, 11, 17-29.
- Comensoli, A., & MacCann, C. (2013). Misconstruing methods and meaning in the General Factor of Personality. *International Journal of Psychology*, 48(4), 625-630.
- Connelly, B. S., & Chang, L. (2016). A Meta-Analytic Multitrait Multirater Separation of Substance and Style in Social Desirability Scales. *Journal of Personality*, 84(3), 319-334.
- Connolly, J. J., Kavanagh, E. J., & Viswesvaran, C. (2007). The convergent validity between self and observer ratings of personality: A meta - analytic review. *International Journal of Selection and Assessment*, 15(1), 110-117.
- Converse, P. D., Peterson, M. H., & Griffith, R. L. (2009). Faking on personality measures: Implications for selection involving multiple predictors. *International Journal of Selection and Assessment*, 17(1), 47-60.

- Conway, J. M., & Lance, C. E. (2010). What reviewers should expect from authors regarding common method bias in organizational research. *Journal of Business and Psychology*, 25(3), 325-334.
- Costa Jr., Paul T. & McCrae, Robert R. (1992). *NEO Personality Inventory Revised* (NEO PI-R). Psychological Assessment Resources Inc. Florida.
- Costa Jr., P. T., & McCrae, R. R. (1995). Domains and facets: Hierarchical personality assessment using the Revised NEO Personality Inventory. *Journal of Personality Assessment*, 64(1), 21-50.
- Cronbach, L. J. (1980). Selection theory for a political world. *Public Personnel Management*, 9(1), 37-50.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281.
- Croson, R., & Sundali, J. (2005). The gambler's fallacy and the hot hand: Empirical data from casinos. *Journal of Risk and Uncertainty*, 30(3), 195-209.
- Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4), 349.
- Danay, E., & Ziegler, M. (2011). Is there really a single factor of personality? A multirater approach to the apex of personality. *Journal of Research in Personality*, 45(6), 560-567.
- Davies, S. E., Connelly, B. S., Ones, D. S., & Birkland, A. S. (2015). The General Factor of Personality: The "Big One," a self-evaluative trait, or a methodological gnat that won't go away? *Personality and Individual Differences*, 81, 13-22.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571-582.

- Depue, R. A., & Collins, P. F. (1999). Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation, and extraversion. *Behavioral and Brain Sciences*, 22(3), 491-517.
- Detert, J. R., Treviño, L. K., & Sweitzer, V. L. (2008). Moral disengagement in ethical decision making: a study of antecedents and outcomes. *Journal of Applied Psychology*, 93(2), 374-391.
- DeYoung, C. G. (2006). Higher-order factors of the Big Five in a multi-informant sample. *Journal of Personality and Social psychology*, 91(6), 1138 - 1151.
- DeYoung, C. G. (2015). Cybernetic big five theory. *Journal of Research in Personality*, 56, 33-58.
- DeYoung, C. G., Peterson, J. B., & Higgins, D. M. (2002). Higher-order factors of the Big Five predict conformity: Are there neuroses of health? *Personality and Individual Differences*, 33(4), 533-552.
- DeYoung, C. G., Quilty, L. C., & Peterson, J. B. (2007). Between facets and domains: 10 aspects of the Big Five. *Journal of Personality and Social Psychology*, 93(5), 880 - 896.
- Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41(1), 417-440.
- Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, 73(6), 1246 - 1256.
- Digman, J. M., & Takemoto-Chock, N. K. (1981). Factors in the natural language of personality: Re-analysis, comparison, and interpretation of six major studies. *Multivariate Behavioral Research*, 16(2), 149-170.
- Dilchert, S., & Ones, D. S. (2012). Application of preventative strategies. *New Perspectives on Faking in Personality Assessment*, 177-200.

- Dilchert, S., Ones, D. S., Viswesvaran, C., & Deller, J. (2006). Response distortion in personality measurement: born to deceive, yet capable of providing valid self-assessments? *Psychology Science*, 48(3), 209 - 225.
- Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance*, 16(1), 81-106.
- Donovan, J. J., Dwight, S. A., & Schneider, D. (2014). The Impact of Applicant Faking on Selection Measures, Hiring Decisions, and Employee Performance. *Journal of Business and Psychology*, 1-15.
- Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical education*, 37(9), 830-837.
- Drasgow, F., Stark, S., Chernyshenko, O. S., Nye, C. D., Hulin, C. L., & White, L. A. (2012). *Development of the Tailored Adaptive Personality Assessment System (TAPAS) to support Army personnel selection and classification decisions*. Drasgow Consulting Group, Urbana, IL.
- Dullaghan, Timothy Ryan (2013). Variance in Faking in High-Stakes Personality Assessment as an Indication of Job Knowledge. *Graduate School Theses and Dissertations*. <http://scholarcommons.usf.edu/etd/4666>
- Duncan, J., Seitz, R. J., Kolodny, J., Bor, D., Herzog, H., Ahmed, A. Newell, F., & Emslie, H. (2000). A neural basis for general intelligence. *Science*, 289(5478), 457-460.
- Dunkel, C. S., van der Linden, D., Brown, N. A., & Mathes, E. W. (2016). Self report based General Factor of Personality as socially-desirable responding, positive self-evaluation, and social-effectiveness. *Personality and Individual Differences*, 92, 143-147.

- Duval, T. S., & Lalwani, N. (1999). Objective self-awareness and causal attributions for self-standard discrepancies: Changing self or changing standards of correctness. *Personality and Social Psychology Bulletin*, 25(10), 1220-1229.
- Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance*, 16(1), 1-23.
- Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behaviour and Human Performance*, 13(2), 171-192.
- Elliot, A. J., & Devine, P. G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology*, 67(3), 382-394.
- Ellingson, J. E. (2012). People fake only when they need to fake. *New Perspectives on Faking in Personality Assessment*, 19-33.
- Ellingson, J. E., Heggstad, E. D., & Makarius, E. E. (2012). Personality retesting for managing intentional distortion. *Journal of Personality and Social Psychology*, 102(5), 1063-1076.
- Ellingson, J. E., Smith, D. B., & Sackett, P. R. (2001). Investigating the influence of social desirability on personality factor structure. *Journal of Applied Psychology*, 86(1), 122-133.
- Embretson (Whitely), S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93(1), 179-197.
- Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher*, 36(8), 449-455.
- Ent, M. R., & Baumeister, R. F. (2014). Obedience, Self - Control, and the Voice of Culture. *Journal of Social Issues*, 70(3), 574-586.
- Equality Act (2004). www.irishstatutebook.ie

- Erdle, S., & Rushton, J. P. (2011). Does self-esteem or social desirability account for a general factor of personality (GFP) in the Big Five? *Personality and Individual Differences*, 50(7), 1152-1154.
- Eysenck, H. J. (1968). Eysenck personality inventory manual. *San Diego: Educational and Industrial Testing Service.*
- Fan, J., Wong, C. C., Carroll, S. A., & Lopez, F. J. (2008). An empirical investigation of the influence of social desirability on the factor structure of the Chinese 16PF. *Personality and Individual Differences*, 45(8), 790-795.
- Fan, J., Gao, D., Carroll, S. A., Lopez, F. J., Tian, T. S., & Meng, H. (2012). Testing the efficacy of a new procedure for reducing faking on personality tests within selection contexts. *Journal of Applied Psychology*, 97(4), 866-880.
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175-191.
- Finch, J. F., & West, S. G. (1997). The investigation of personality structure: Statistical models. *Journal of Research in Personality*, 31(4), 439-485.
- Fischbacher, U. & Heusi, F. 2008. "Lies in Disguise. An experimental study on cheating". TWI Research Paper Series.
- Fischbacher, U., & Föllmi - Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, 11(3), 525-547.
- Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: biased assessments of others' ability to read one's emotional states. *Journal of personality and social psychology*, 75(2), 332.

- Gnambs, T. (2013). The Elusive General Factor of Personality: The Acquaintance Effect. *European Journal of Personality*, 27(5), 507-520.
- Goodwin, L. D., & Leech, N. L. (2003). The meaning of validity in the new standards for educational and psychological testing: Implications for measurement courses. *Measurement and evaluation in Counseling and Development*, 36(3), 181-192.
- Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48, 26-26.
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1), 84-96.
- Gottfredson Linda S. (1997). Why 'g' matters: The complexity of everyday life. *Intelligence*, 24(1), 79-132.
- Griffith, R. L., & Converse, P. D. (2012). The rules of evidence and the prevalence of applicant faking. *New Perspectives on Faking in Personality Assessment*, 1, 34-52.
- Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behaviour. *Personnel Review*, 36(3), 341-355.
- Griffith, R. L., & Peterson, M. H. (2008). The failure of social desirability measures to capture applicant faking behaviour. *Industrial and Organizational Psychology*, 1(3), 308-311.
- Griffith, R. L., & Peterson, M. H. (2011). One piece at a time: the puzzle of applicant faking and a call for theory. *Human Performance*, 24(4), 291-301.

- Grijalva, E., Harms, P. D., Newman, D. A., Gaddis, B. H., & Fraley, R. C. (2013). Narcissism and Leadership: a meta - analytic review of linear and nonlinear relationships. *Personnel Psychology*, 68(1), 1-47.
- Guastello, Stephen J. (1993) A two-(and-a-half)-tiered trait taxonomy. *American Psychologist*, 48(12), 1298-1299.
- Haier, R. J., Colom, R., Schroeder, D. H., Condon, C. A., Tang, C., Eaves, E., & Head, K. (2009). Gray matter and intelligence factors: Is there a neuro-g? *Intelligence*, 37(2), 136-144.
- Hall, R. C. W., & Hall, R. C. W. (2012). Plaintiffs who malingering: impact of litigation on fake testimony. *New Perspectives on Faking in Personality Assessment*, 255-281.
- Hathaway, S. R., McKinley, J. C. (1989). *MMPI-2: Minnesota Multiphasic Personality Inventory-2: manual for administration and scoring*. University of Minnesota Press.
- Hauenstein, N. M., Bradley, K. M., O'Shea, P. G., Shah, Y. J., & Magill, D. P. (2017). Interactions between motivation to fake and personality item characteristics: Clarifying the process. *Organizational Behavior and Human Decision Processes*, 138, 74-92.
- Hausknecht, J. P. (2010). Candidate persistence and personality test practice effects: Implications for staffing system management. *Personnel Psychology*, 63(2), 299-324.
- Heim A.W. (1970). The AH4 group test of intelligence. Windsor: NFER-Nelson.

- Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, 66(3), 393-416.
- Hiemer, J., & Abele, A. E. (2012). High power= Motivation? Low power= Situation? The impact of power, power stability and power motivation on risk-taking. *Personality and Individual Differences*, 53(4), 486-490.
- Higgins, E. T. (1987). Self-discrepancy: a theory relating self and affect. *Psychological Review*, 94(3), 319-340.
- Highhouse, S. (1998). Understanding and improving job-finalist choice: The relevance of behavioral decision research. *Human Resource Management Review*, 7(4), 449-470.
- Highhouse, S. (2002). Assessing the candidate as a whole: A historical and critical analysis of individual psychological assessment for personnel decision making. *Personnel Psychology*, 55(2), 363-396.
- Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology*, 1(3), 333-342.
- Hirsh, J. B., DeYoung, C. G., & Peterson, J. B. (2009). Meta-traits of the Big Five differentially predict engagement and restraint of behaviour. *Journal of Personality*, 77(4), 1085-1102.
- Hoffman, E. (2000). *Ace the Corporate Personality Test*. McGraw Hill Professional.
- Hoffman, L. (2017). Retrived from http://www.lesahoffman.com/CLP948/CLP948_Lecture04_CFA.pdf . January 2017.

- Hofstee, W. K., De Raad, B., & Goldberg, L. R. (1992). Integration of the Big Five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, 63(1), 146-163.
- Hogan, R. (2005). In defense of personality measurement: New wine for old whiners. *Human Performance*, 18(4), 331-341.
- Hogan, J., & Holland, B. (2003). Using theory to evaluate personality and job-performance relations: a socioanalytic perspective. *Journal of Applied Psychology*, 88(1), 100 - 112.
- Hogan, R., & Hogan, J. (1995). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Assessment Systems.
- Hogan, R., & Hogan, J. (2001). Assessing leadership: A view from the dark side. *International Journal of Selection and Assessment*, 9(1 - 2), 40-51.
- Hogan, R., Hogan, J., & Roberts, B. W. (1996). Personality measurement and employment decisions: Questions and answers. *American psychologist*, 51(5), 469-477.
- Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology*, 92(5), 1270-1285.
- Hogan, J., Hogan, R., & Kaiser, R. B. (2010). Management derailment. *American Psychological Association Handbook of Industrial and Organizational Psychology*, 3, 555-575.
- Holden, R. R. (2008). Underestimating the effects of faking on the validity of self report personality scales. *Personality and Individual Differences*, 44(1), 311-321.
- Holden, R. R., & Book, A. S. (2012). Faking does distort self report personality assessment. *New Perspectives on Faking in Personality Assessment*, 71-84.

- Holladay, C. L., David, E., & Johnson, S. K. (2013). Retesting personality in employee selection: implications of the context, sample, and setting. *Psychological Reports, 112*(2), 486-501.
- Hollenbeck, G. P. (2009). Executive selection—What's right... and what's wrong? *Industrial and Organizational Psychology, 2*(2), 130-143.
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Journal of Business Research Methods, 61*(1), 53-60.
- Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review, 14*(3), 332-346.
- Hopwood, C. J., Wright, A. G., & Donnellan, B. M. (2011). Evaluating the evidence for the general factor of personality across multiple inventories. *Journal of Research in Personality, 45*(5), 468-478.
- Hough, L. M. (1992). The 'Big Five' personality variables--construct confusion: Description versus prediction. *Human Performance, 5*(1-2), 139-155.
- Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance, 11*(2-3), 209-244.
- Hough, L. M., & Oswald, F. L. (2008). Personality testing and industrial—organizational psychology: Reflections, progress, and prospects. *Industrial and Organizational Psychology, 1*(3), 272-290.
- Hough, L. M., Ones, D. S., & Viswesvaran, C. (1998). Personality correlates of managerial performance constructs. In *13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas*.

- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55.
- Hughes D.J. (in press). Psychometric Validity: Establishing the accuracy and appropriateness of psychometric measures. In Irwing, P., Booth, T. & Hughes, D.J. *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Approach to Survey, Scale and Test Development*. Chichester: Wiley.
- Hughes, D. J., & Batey, M. (in press). Using personality questionnaires for selection. In H. Goldstein, E. Pulakos, J. Passmore, & C. Semedo (Eds.). *The Wiley Handbook of the Psychology of Recruitment, Selection & Retention*. Chichester: Wiley-Blackwell.
- Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, 91(3), 594-612.
- Huselid, M. A. (1995). The impact of human resource management practices on turnover, productivity, and corporate financial performance. *Academy of management journal*, 38(3), 635-672.
- Ioannidis, J. P. (2005). Why most published research findings are false. *Chance*, 18(4), 40-47.
- Ion, A., & Iliescu, D. (2017). The measurement equivalence of personality measures across high-and low-stake test taking settings. *Personality and Individual Differences*, 110, 1-6.
- Institute of Psychometric Coaching (2017). Retrived from <http://www.psychometricinstitute.com.au/Personal-Psychometric-Coaching.html> . January 2017.

- Jacobsen, C., Fosgaard, T. R., & Pascual - Ezama, D. (2017). Why do we lie? a practical guide to the dishonesty literature. *Journal of Economic Surveys*. In press.
- John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of Personality: Theory and Research*, 2, 102-138.
- John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative big five trait taxonomy. *Handbook of Personality: Theory and Research*, 3, 114-158.
- John, O. P., Donahue, E. M., & Kentle, R. L. (1991). The Big Five Inventory—Versions 4a and 54 (Tech. Rep.). Berkeley: Institute of Personality and Social Research, University of California.
- Johnson, J. A. (1994). Clarification of factor five with the help of the AB5C model. *European Journal of Personality*, 8(4), 311-334.
- Johnson, J. A., & Ostendorf, F. (1993). Clarification of the five-factor model with the Abridged Big Five Dimensional Circumplex. *Journal of Personality and Social Psychology*, 65, 563-563.
- Johnson, R. E., Rosen, C. C., & Chang, C. H. (2011a). To aggregate or not to aggregate: Steps for developing and validating higher-order multidimensional constructs. *Journal of Business and Psychology*, 26(3), 241-248.
- Johnson, R. E., Rosen, C. C., & Djurdjevic, E. (2011b). Assessing the impact of CMV on higher order multidimensional constructs. *Journal of Applied Psychology*, 96(4), 744-761.
- Johnson, R. E., Rosen, C. C., Chang, C. H. D., Djurdjevic, E., & Taing, M. U. (2012). Recommendations for improving the construct clarity of higher-order

- multidimensional constructs. *Human Resource Management Review*, 22(2), 62-72.
- Jonason, P. K., Slomski, S., & Partyka, J. (2012). The Dark Triad at work: How toxic employees get their way. *Personality and Individual Differences*, 52(3), 449-453.
- Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The big five personality traits, general mental ability, and career success across the life span. *Personnel Psychology*, 52(3), 621-652.
- Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: a qualitative and quantitative review. *Journal of Applied Psychology*, 87(4), 765-780.
- Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2003). The core self - evaluations scale: Development of a measure. *Personnel Psychology*, 56(2), 303-331.
- Judge, T. A., Rodell, J. B., Klinger, R. L., Simon, L. S., & Crawford, E. R. (2013). Hierarchical representations of the five-factor model of personality in predicting job performance: Integrating three organizing frameworks with two theoretical perspectives. *Journal of Applied Psychology*, 98, 875-925.
- Just, C. (2011). A review of literature on the general factor of personality. *Personality and Individual Differences*, 50(6), 765-771.
- Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38(4), 319-342.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73.

- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263-291.
- Kandler, C., Riemann, R., Spinath, F. M., & Angleitner, A. (2010). Sources of variance in personality facets: A multiple-rater twin study of self-peer, peer-peer, and self-self (dis)agreement. *Journal of Personality*, 78, 1565–1594.
- Keltner, D., Gruenfeld, D. H., & Anderson, C. (2003). Power, approach, and inhibition. *Psychological Review*, 110(2), 265-284.
- Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. *The Handbook of Social Psychology*, 1(4), 233-265. Updated at http://davidakenny.net/cm/identify_formal.htm
- King, L. A., Walker, L. M., & Broyles, S. J. (1996). Creativity and the five-factor model. *Journal of research in personality*, 30(2), 189-203.
- Klehe, U. C., Kleinmann, M., Hartstein, T., Melchers, K. G., König, C. J., Heslin, P. A., & Lievens, F. (2012). Responding to personality tests in a selection context: The role of the ability to identify criteria and the ideal-employee factor. *Human Performance*, 25(4), 273-302.
- Kleinmann, M., Ingold, P. V., Lievens, F., Jansen, A., Melchers, K. G., & König, C. J. (2011). A different look at why selection procedures work. The role of candidates' ability to identify criteria. *Organizational Psychology Review*, 1(2), 128-146.
- Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling*. 4th Ed. New York: Guilford Press.
- Kline, R. B. (2013). Exploratory and confirmatory factor analysis. In Y. Petscher & C. Schatschneider (Eds.), *Applied Quantitative Analysis in the Social Sciences* (pp. 171-207). New York: Routledge.

- Kluger, A. N., & Colella, A. (1993). Beyond the mean bias: The effect of warning against faking on biodata item variances. *Personnel Psychology*, 46(4), 763-780.
- Komar, S., Brown, D. J., Komar, J. A., & Robie, C. (2008). Faking and the validity of conscientiousness: A Monte Carlo investigation. *Journal of Applied Psychology*, 93(1), 140-154.
- Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking “big” personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. *Psychological Bulletin*, 136(5), 768-821.
- Kuncel, N. R., Borneman, M., & Kiger, T. (2011). Innovative item response process and Bayesian faking detection methods: More questions than answers. *New perspectives on faking in personality assessment*, 102-112.
- Kurtz, J. E., Tarquini, S. J., & Iobst, E. A. (2008). Socially desirable responding in personality assessment: Still more substance than style. *Personality and Individual Differences*, 45(1), 22-27.
- Lance, C. E., & Jackson, D. J. (2015). Seek and ye shall find. *Industrial and Organizational Psychology*, 8(03), 452-463.
- Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The Sources of Four Commonly Reported Cut-off Criteria What Did They Really Say? *Organizational Research Methods*, 9(2), 202-220.
- Lance, C. E., Dawson, B., Birkelbach, D., & Hoffman, B. J. (2010). Method effects, measurement error, and substantive conclusions. *Organizational Research Methods*, 13(3), 435-455.

- Landers, R. N., Sackett, P. R., & Tuzinski, K. A. (2011). Retesting after initial failure, coaching rumors, and warnings against faking in online personality measures for selection. *Journal of Applied Psychology, 96*(1), 202 - 210.
- Landis, R. S., Edwards, B. D., & Cortina, J. M. (2009). On the practice of allowing correlated residuals among indicators in structural equation models. *Statistical and methodological myths and urban legends: Doctrine, verity, and fable in the organizational and social sciences*, 195-214.
- LaPiere, R. T. (1934). Attitudes vs. actions. *Social Forces, 13*(2), 230-237.
- Le, H., Schmidt, F. L., & Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods, 12*(1), 165-200.
- Lease, D. R. (2006). From Great to Ghastly: How Toxic Organizational Cultures Poison Companies The Rise and Fall of Enron, WorldCom, HealthSouth, and Tyco International. *Academy of Business Education, April*, 6-7.
- Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment, 14*(2), 131-141.
- Lim, B. C., & Ployhart, R. E. (2006). Assessing the Convergent and Discriminant Validity of Goldberg's International Personality Item Pool A Multitrait-Multimethod Examination. *Organizational Research Methods, 9*(1), 29-54.
- Lindell, M. K., & Whitney, D. J. (2001). Accounting for CMV in cross-sectional research designs. *Journal of Applied Psychology, 86*(1), 114-121.
- Loehlin, J. C., & Martin, N. G. (2011). The general factor of personality: Questions and elaborations. *Journal of Research in Personality, 45*(1), 44-49.

- Lönnqvist, J. E., Irlenbusch, B., & Walkowitz, G. (2014). Moral hypocrisy: impression management or self-deception? *Journal of Experimental Social Psychology*, 55, 53-62.
- MacCann, C., Ziegler, M., & Roberts, R. D. (2011). Faking in personality assessment – reflections and recommendations. *New Perspectives on Faking in Personality Assessment*, 309-329.
- Malhotra, N. K., Kim, S. S., & Patil, A. (2006). Common method variance in IS research: A comparison of alternative approaches and a reanalysis of past research. *Management Science*, 52(12), 1865-1883.
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3), 519-530.
- Markon, K. E., Krueger, R. F., & Watson, D. (2005). Delineating the structure of normal and abnormal personality: an integrative hierarchical approach. *Journal of Personality and Social Psychology*, 88(1), 139-157.
- Marcus, B. (2006). Relationships between faking, validity, and decision criteria in personnel selection. *Psychology Science*, 48(3), 226-246.
- Marcus, B. (2009). 'Faking' From the Applicant's Perspective: A theory of self - presentation in personnel selection settings. *International Journal of Selection and Assessment*, 17(4), 417-430.
- Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, 33(2), 181-220.
- Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and

- confirmatory factor analysis. *Annual Review of Clinical Psychology*, 10, 85-110.
- Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological assessment*, 22(3), 471.
- Maslow, A. H. (1948). "Higher" and "Lower" Needs. *The Journal of Psychology*, 25(2), 433-436.
- Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, 45(6), 633-644.
- McCrae, R. R. (2010). The place of the FFM in personality psychology. *Psychological Inquiry*, 21(1), 57-64.
- McCrae, R. R., & Costa Jr, P. T. (1999). A five-factor theory of personality. *Handbook of Personality: Theory and Research*, 2, 139-153.
- McCrae, R. R., & Costa, P. T. (2008). Empirical and theoretical status of the five-factor model of personality traits. *The SAGE handbook of personality theory and assessment*, 1, 273-294.
- McCrae, R. R., & Terracciano, A. (2005). Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology*, 88(3), 547-561.
- McCrae, R. R., Costa, Jr, P. T., & Martin, T. A. (2005). The NEO-PI-3: A more readable revised NEO personality inventory. *Journal of personality assessment*, 84(3), 261-270.
- McCrae, R. R., Zonderman, A. B., Costa, P. T., Bond, M. H., & Purnonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality

- inventory: confirmatory factor analysis versus procrustes rotation. *Journal of Personality and Social Psychology*, 70, 552-566.
- McCrae, R. R., Yamagata, S., Jang, K. L., Riemann, R., Ando, J., Ono, Y., & Spinath, F. M. (2008). Substance and artifact in the higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, 95(2), 442-455.
- McDonald, R.P. (1999). *Test Theory – A Unified Treatment*. Laurence Erlbaum Associates, New Jersey.
- McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, 7(1), 64.
- McFarland, L. A. (2003). Warning against faking on a personality test: Effects on applicant reactions and personality test scores. *International Journal of Selection and Assessment*, 11(4), 265-276.
- McFarland, L. A., Ryan, A. M., & Ellis, A. (2002). Item placement on a personality measure: Effects on faking behaviour and test measurement properties. *Journal of Personality Assessment*, 78(2), 348-369.
- MacKenzie, S. B. (2003). The dangers of poor construct conceptualization. *Journal of the Academy of Marketing Science*, 31(3), 323-326.
- MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly*, 35(2), 293-334.
- Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology*, 77, 531-551.

- Meeus, W. H., & Raaijmakers, Q. A. (1986). Administrative Obedience: Carrying Out Orders to Use Psychological - Administrative Violence. *European Journal of Social Psychology*, 16(4), 311-324.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5-11.
- Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749.
- Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, 45(1-3), 35-44.
- Milgram, S. (1963). Behavioral Study of obedience. *The Journal of Abnormal and Social Psychology*, 67(4), 371.
- Milligan, G. W., & Cooper, M. C. (1987). Methodology review: Clustering methods. *Applied Psychological Measurement*, 11(4), 329-354.
- Mischel, W. (1969). Continuity and change in personality. *American Psychologist*, 24(11), 1012-1018.
- Mlodinow, L. (2009). *The drunkard's walk: How randomness rules our lives*. Vintage.
- Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, 60(3), 683-729.
- Morin, A. (2011). Self - awareness part 1: Definition, measures, effects, functions, and antecedents. *Social and Personality Psychology Compass*, 5(10), 807-823.
- Möttus R, Kandler C, Bleidorn W, Riemann R, McCrae RR. (2017). Personality traits below facets: the consensual validity, longitudinal stability, Heritability, and

- utility of personality nuances. *Journal of Personality and Social Psychology*, 112(3), 474-490.
- Mount, M. K., Barrick, M. R., & Callans, M. (1995). Manual for the Personal Characteristics Inventory. *Libertyville, IL: Wonderlic Personnel Test, Inc.*
- Mount, M. K., Barrick, M. R., & Stewart, G. L. (1998). Five-factor model of personality and performance in jobs involving interpersonal interactions. *Human performance*, 11(2-3), 145-165.
- Moutafi, J., Furnham, A., & Crump, J. (2006). What facets of openness and conscientiousness predict fluid intelligence score?. *Learning and Individual Differences*, 16(1), 31-42.
- Mueller-Hanson, R., Heggstad, E. D., & Thornton III, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology*, 88(2), 348-355.
- Murphy, K. R. (2005). Why don't measures of broad dimensions of personality perform better as predictors of job performance? *Human Performance*, 18(4), 343-357.
- Murphy, K. R., & Shiarella, A. H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests: Multivariate frameworks for studying test validity. *Personnel Psychology*, 50(4), 823-854.
- Murphy, K. & Davidshofer, C. (1998). *Psychological Testing – Principles and Applications*. 4th Ed., Prentice Hall, New Jersey.
- Musek, J. (2007). A general factor of personality: Evidence for the Big One in the five-factor model. *Journal of Research in Personality*, 41(6), 1213-1233.
- Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied Linear Statistical Models*. Chicago: Irwin

- Newton, P.E. & Baird, J. (2016). The great validity debate. *Assessment in Education: Principles, Policy & Practice*, 23:2, 173-177,
- O'Brien, R. M. (1994). Identification of simple measurement models with multiple latent variables and correlated errors. *Sociological Methodology*, 24, 137-170.
- Oh, I. S., Kim, S., & Van Iddekinge, C. H. (2015). Taking it to another level: Do personality-based human capital resources matter to firm performance? *Journal of Applied Psychology*, 100(3), 935-947.
- Ones, D. S., & Dilchert, S. (2009). How special are executives? How special should executive selection be? Observations and recommendations. *Industrial and Organizational Psychology*, 2(2), 163-170.
- Ones, D. S., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. *Human Performance*, 11(2-3), 245-269.
- Ones, D. S., & Viswesvaran, C. (2001). Integrity Tests and Other Criterion Focused Occupational Personality Scales (COPS) Used in Personnel Selection. *International Journal of Selection and Assessment*, 9(1-2), 31-39.
- Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, 81(6), 660-679.
- Ones, D.S., Viswesvaran, C., Dilchert S., (2005). Personality at work: Raising awareness and correcting misconceptions. *Human Performance*, 18, 389–404.
- Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology*, 60(4), 995-1027.

- Osborne, J. W. (2014). *Best Practices in Exploratory Factor Analysis*. Scotts Valley, CA: CreateSpace Independent Publishing.
- Osborne, J. W., & Costello, A. B. (2004). Sample size and subject to item ratio in principal components analysis. *Practical Assessment, Research & Evaluation*, 9(11), 8.
- Ozer, D. J. (1999). Four principles for personality assessment, in Pervin, L. A., & John, O. P. (1999). *Handbook of Personality: Theory and Research*. Elsevier.
- Pace, V. L., & Borman, W. C. (2006). The use of warning to discourage faking on noncognitive inventories. *A Closer Examination of Applicant Faking Behavior*, 283-304, IAP.
- Padilla, A., Hogan, R., & Kaiser, R. B. (2007). The toxic triangle: Destructive leaders, susceptible followers, and conducive environments. *The Leadership Quarterly*, 18(3), 176-194.
- Paul, A.M. (2004). *The Cult of Personality Testing: How Personality Tests Are Leading Us to Miseducate Our Children, Mismanage Our Companies, and Misunderstand Ourselves*. Free Press, New York
- Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46(3), 598-609.
- Paulhus, D. L. (1998). Paulhus deception scales (PDS): *The balanced inventory of desirable responding – 7*. North Tonawanda, NY: Multi-Health Systems.
- Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. *The Role of Constructs in Psychological and Educational Measurement*, 49-69. Routledge
- Paulhus, D. L., & Vazire, S. (2007). The self-report method. *Handbook of research methods in personality psychology*, 1, 224-239.

- Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, 60(2), 307-317.
- Paulhus, D. L., & Williams, K. M. (2002). The dark triad of personality: narcissism, machiavellianism, and psychopathy. *Journal of Research in Personality*, 36(6), 556-563.
- Paulhus, D. L., Harms, P. D., Bruce, N.M., & Lysy, D. C. (2003). The Over-Claiming Technique: Measuring Self-Enhancement Independent of Ability. *Journal of Personality and Social Psychology*, 84(4), 890–904.
- Paunonen, S. V., & LeBel, E. P. (2012). Socially desirable responding and its elusive effects on the validity of personality assessments. *Journal of Personality and Social Psychology*, 103(1), 158-170.
- Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, 8(2), 287-312.
- Perinelli, E., & Gremigni, P. (2016). Use of social desirability scales in clinical psychology: a systematic review. *Journal of clinical psychology*, 72(6), 534-551.
- Peterson, M. H., Griffith, R. L., & Converse, P. D. (2009). Examining the role of applicant faking in hiring decisions: Percentage of fakers hired and hiring discrepancies in single-and multiple-predictor selection. *Journal of Business and Psychology*, 24(4), 373-386.
- Pittenger, D. J. (1993). Measuring the MBTI... and coming up short. *Journal of Career Planning and Employment*, 54(1), 48-52.
- Pittenger, D. J. (2005). Cautionary comments regarding the Myers-Briggs Type Indicator. *Consulting Psychology Journal: Practice and Research*, 57(3), 210.

- Pittarello, A., Leib, M., Gordon-Hecker, T., & Shalvi, S. (2015). Justifications shape ethical blind spots. *Psychological Science*, 26(6), 794-804.
- Ployhart, R. E., Schmitt, N., & Tippins, N. T. (2017). Solving the Supreme Problem: 100 Years of Selection and Recruitment at the Journal of Applied Psychology. *Journal of Applied Psychology*, 102(3), 1-14.
- Popper, K. (1963). *Conjectures and refutations: The growth of scientific knowledge*. Routledge.
- Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, 63, 539-569.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879-903.
- Prabhakaran, V., Smith, J. A., Desmond, J. E., Glover, G. H., & Gabrieli, J. D. (1997). Neural substrates of fluid reasoning: an fMRI study of neocortical activation during performance of the Raven's Progressive Matrices Test. *Cognitive psychology*, 33(1), 43-63.
- Pruckner, G. J., & Sausgruber, R. (2013). Honesty on the streets: A field study on newspaper purchasing. *Journal of the European Economic Association*, 11(3), 661-679.
- Pryor, J. B., Gibbons, F. X., Wicklund, R. A., Fazio, R. H., & Hood, R. (1977). Self-focused attention and self report validity. *Journal of Personality*, 45(4), 513-527.
- Quinn, T. J. (1999). Practical realization of the definition of the metre (1997). *Metrologia*, 36(3), 211.

- Raychaudhuri, S. (2008). Introduction to monte carlo simulation. In *Simulation Conference, 2008. WSC 2008. Winter* (pp. 91-100). IEEE.
- Raven, J. C. (1965). *Advanced Progressive Matrices: Sets I and II*. London: Lewis.
- Reilly, T., & O'Brien, R. M. (1996). Identification of Confirmatory Factor Analysis Models of Arbitrary Complexity The Side-by-Side Rule. *Sociological Methods and Research*, 24(4), 473-491.
- Revelle, W., & Wilt, J. (2013). The general factor of personality: A general critique. *Journal of research in personality*, 47(5), 493-504.
- Resick, C. J., Whitman, D. S., Weingarden, S. M., & Hiller, N. J. (2009). The bright-side and the dark-side of CEO personality: Examining core self-evaluations, narcissism, transformational leadership, and strategic influence. *Journal of Applied Psychology*, 94(6), 1365-1381.
- Riemann, R., & Kandler, C. (2010). Construct validation using multitrait - multimethod twin data: The case of a general factor of personality. *European Journal of Personality*, 24(3), 258-277.
- Richardson, H., Simmering, M., & Sturman, M. (2009). A tale of three perspectives: Examining post hoc statistical techniques for detection and correction of CMV. *Organizational Research Methods*, 12, 762–800.
- Roberts, B. W. (2009). Back to the future: Personality and Assessment and personality development. *Journal of Research in Personality*, 43(2), 137-145.
- Roberts, B. W., & Caspi, A. (2001). Personality Development and the Person-Situation Debate: It's Déjà Vu All Over Again. *Psychological Inquiry*, 12(2), 104-109.
- Roberts, B. W., & Jackson, J. J. (2008). Sociogenomic personality psychology. *Journal of Personality*, 76(6), 1523-1544.

- Roberts, B. W., Caspi, A., & Moffitt, T. E. (2003). Work experiences and personality development in young adulthood. *Journal of Personality and Social psychology*, 84(3), 582-593.
- Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological Bulletin*, 132(1), 1-25.
- Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A., & Goldberg, L. R. (2007). The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science*, 2(4), 313-345.
- Robie, C., Brown, D. J., & Beaty, J. C. (2007). Do people fake on personality inventories? A verbal protocol analysis. *Journal of Business and Psychology*, 21(4), 489-509.
- Robie, C., Komar, S., & Brown, D. J. (2010). The effects of coaching and speeding on Big Five and impression management scale scores. *Human Performance*, 23(5), 446-467.
- Rodgers, J. L. (2010). The Epistemology of Mathematical and Statistical Modeling - A Quiet Methodological Revolution. *American Psychologist*, 65(1), 1-12.
- Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of Consulting Psychology*, 21(2), 95-103.
- Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology*, 83(4), 634-644.

- Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: what does current research support? *Human Resource Management Review*, 16(2), 155-180.
- Ruedy, N. E., Moore, C., Gino, F., & Schweitzer, M. E. (2013). The cheater's high: The unexpected affective benefits of unethical behavior. *Journal of Personality and Social Psychology*, 105 (4), 531–548.
- Ruiz, J. M., Smith, T. W., & Rhodewalt, F. (2001). Distinguishing narcissism and hostility: Similarities and differences in interpersonal circumplex and five-factor correlates. *Journal of Personality Assessment*, 76(3), 537-555.
- Rushton, J. P. (1985). Differential K theory: The sociobiology of individual and group differences. *Personality and Individual Differences*, 6(4), 441-452.
- Rushton, J. P., & Irwing, P. (2008). A General Factor of Personality (GFP) from two meta-analyses of the Big Five. *Personality and Individual Differences*, 45(7), 679-683.
- Rushton, J. P., Bons, T. A., & Hur, Y. M. (2008). The genetics and evolution of the general factor of personality. *Journal of Research in Personality*, 42(5), 1173-1185.
- Rushton, J. P., Bons, T. A., Ando, J., Hur, Y. M., Irwing, P., Vernon, P. A., & Barbaranelli, C. (2009). A general factor of personality from multitrait-multimethod data and cross-national twins. *Twin Research and Human Genetics*, 12(4), 356-365.
- Rynes, S. L., Colbert, A. E., & Brown, K. G. (2002). HR professionals' beliefs about effective human resource practices: Correspondence between research and practice. *Human Resource Management*, 41, 149– 174.

- Sackett, P. R. (2011). Integrating and prioritizing theoretical perspectives on applicant faking of personality measures. *Human Performance*, 24(4), 379-385.
- Sackett, P. R.. (2012). Faking in personality assessment – where do we stand? *New Perspectives on Faking in Personality Assessment*, 330-344.
- Sackett, P. R., Lievens, F., Berry, C. M., & Landers, R. N. (2007). A cautionary note on the effects of range restriction on predictor intercorrelations. *Journal of Applied Psychology*, 92(2), 538-544.
- Saggino, A., & Kline, P. (1996). Item factor analysis of the seventy-one experimental items of the Italian version of the Myers-Briggs Type Indicator. *Personality and Individual Differences*, 21(3), 441-444.
- Saggino, A., Cooper, C., & Kline, P. (2001). A confirmatory factor analysis of the Myers–Briggs Type Indicator. *Personality and Individual Differences*, 30(1), 3-9.
- Salgado, J. F. (2003). Predicting job performance using FFM and non FFM personality measures. *Journal of Occupational and Organizational Psychology*, 76(3), 323-346.
- Salgado, J. F. (2005). Personality and social desirability in organizational settings: practical implications for work and organizational psychology. *Papeles del Psicólogo*, 26(S 124).
- Salgado, J. F. (2016). A Theoretical Model of Psychometric Effects of Faking on Assessment Procedures: Empirical findings and implications for personality at work. *International Journal of Selection and Assessment*, 24(3), 209-228.
- Salgado, J. F., Moscoso, S., & Lado, M. (2003). Evidence of cross-cultural invariance of the big five personality dimensions in work settings. *European Journal of Personality*, 17(S1), S67-S76.

- Samuel, D. B., & Widiger, T. A. (2008). A meta-analytic review of the relationships between the five-factor model and DSM-IV-TR personality disorders: A facet level analysis. *Clinical Psychology Review*, 28(8), 1326 - 1342.
- Santos, J. M., & Horta, H. (2015). The generational gap of science: a dynamic cluster analysis of doctorates in an evolving scientific system. *Scientometrics*, 104(1), 381-406.
- Saucier, G. (2002). Orthogonal markers for orthogonal factors: The case of the Big Five. *Journal of Research in Personality*, 36(1), 1-31.
- Saucier, G., & Goldberg, L. R. (1998). Assessing the Big Five: Applications of 10 psychometric criteria to the development of marker scales. In *Big Five Assessment*, 29-58, de Raad, B. E., & Perugini, M. E. (Eds.). Hogrefe & Huber Publishers.
- Schermer, J. A., & Vernon, P. A. (2010). The correlation between general intelligence (g), a general factor of personality (GFP), and social desirability. *Personality and Individual Differences*, 48(2), 187-189.
- Schmidt, Frank L., & Hunter, John (2004) General Mental Ability in the World of Work: Occupational Attainment and Job Performance. *Journal of Personality and Social Psychology*, 86(1), 162-173.
- Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology*, 78(6), 966-974.
- Schmitt, N., & Oswald, F. L. (2006). The impact of corrections for faking on the validity of noncognitive measures in selection settings. *Journal of Applied Psychology*, 91(3), 613-621.

- Schumm, W. R. (2010). Statistical requirements for properly investigating a null hypothesis. *Psychological Reports*, 107(3), 953-971.
- Shaver, P. R., Brennan, K. A., Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1991). Measures of personality and social psychological attitudes. *Robinson JP, Shaver PR, Wrightsman LS, eds. Measures of Depression and Loneliness, 1*, 212-215.
- Shu, L. L., Gino, F., & Bazerman, M. H. (2011). Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. *Personality and Social Psychology Bulletin*, 37(3), 330-349.
- Shu, L. L., Mazar, N., Gino, F., Ariely, D., & Bazerman, M. H. (2012). Signing at the beginning makes ethics salient and decreases dishonest self reports in comparison to signing at the end. *Proceedings of the National Academy of Sciences*, 109(38), 15197-15200.
- Silvia, P. J., & Duval, T. S. (2001). Objective self-awareness theory: Recent progress and enduring problems. *Personality and Social Psychology Review*, 5, 230-241.
- Singh, J. (2008). Impostors Masquerading as Leaders: Can the Contagion be Contained? *Journal of Business Ethics*, 82(3), 733-745.
- Sjöberg, L. (2015). Correction for faking in self - report personality tests. *Scandinavian Journal of Psychology*, 56(5), 582-591.
- Sliter, K. A., & Christiansen, N. D. (2012). Effects of targeted self-coaching on applicant distortion of personality measures. *Journal of Personnel Psychology*, 11(4), 169-175.
- Smith, G. T. (2005). On construct validity: issues of method and measurement. *Psychological Assessment*, 17(4), 396.

- Smith, D. B., & Ellingson, J. E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology*, 87(2), 211 - 219.
- Smith, D. B., Hanges, P. J., & Dickson, M. W. (2001). Personnel selection and the five-factor model: Reexamining the effects of applicant's frame of reference. *Journal of Applied Psychology*, 86(2), 304-315.
- Sollman, M. J., & Berry, D. T. (2011). Detection of inadequate effort on neuropsychological testing: A meta-analytic update and extension. *Archives of Clinical Neuropsychology*, 26, 774–789.
- Spector, P. E. (2006). Method variance in organizational research truth or urban legend? *Organizational Research Methods*, 9(2), 221-232.
- Spector, P. E., Rosen, C. C., Richardson, H. A., Williams, L. J., & Johnson, R. E. (2017). A New Perspective on Method Variance: A Measure-Centric Approach. *Journal of Management*. In press.
- Spence, I. (1983). Monte Carlo simulation studies. *Applied Psychological Measurement*, 7(4), 405-425.
- Stark, S., Chernyshenko, O. S., Chan, K. Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology*, 86(5), 943-953.
- Stein, M. (2003). Unbounded irrationality: Risk and organizational narcissism at long term capital management. *Human Relations*, 56(5), 523-540.
- Stevens, G. W., Deuling, J. K., & Armenakis, A. A. (2012). Successful psychopaths: Are they unethical decision-makers and why? *Journal of Business Ethics*, 105(2), 139-149.

- Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, 5, 1-25.
- Sun, J., Kaufman, S. B., & Smillie, L. D. (2016). Unique associations between big five personality aspects and multiple dimensions of well - being. *Journal of Personality* (in press).
- Templer, D. I. (2013). Rushton: The great theoretician and his contribution to personality. *Personality and Individual Differences*, 55(3), 243-246.
- Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology*, 44(4), 703-742.
- Tett, R. P., & Christiansen, N. D. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007). *Personnel Psychology*, 60(4), 967-993.
- Thiele, S., Kubacki, K., Tkaczynski, A., & Parkinson, J. (2015). Using two-step cluster analysis to identify homogeneous physical activity groups. *Marketing Intelligence & Planning*, 33(4), 522 -537.
- Tracey, T. J. (2016). A note on socially desirable responding. *Journal of Counseling Psychology*, 63(2), 224.
- Tupes, E. C., & Christal, R. E. (1992). Recurrent personality factors based on trait ratings. *Journal of Personality*, 60(2), 225-251.
- Twenge, J. M. (2009). Status and gender: The paradox of progress in an age of narcissism. *Sex Roles*, 61(5-6), 338-340.
- Uziel, L. (2010). Rethinking social desirability scales from impression management to interpersonally oriented self-control. *Perspectives on Psychological Science*, 5(3), 243-262.

- Uziel, L. (2014). Impression Management (“Lie”) Scales Are Associated With Interpersonally Oriented Self - Control, Not Other - Deception. *Journal of Personality*, 82(3), 200-212.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4-70.
- Van der Linden, D., Bakker, A. B., & Serlie, A. W. (2011). The General Factor of Personality in selection and assessment samples. *Personality and Individual Differences*, 51(5), 641-645.
- Van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The general factor of personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. *Journal of Research in Personality*, 44(3), 315-327.
- Van der Linden, D., Vreeke, L., & Muris, P. (2013). Don’t be afraid of the General Factor of Personality (GFP): Its relationship with behavioral inhibition and anxiety symptoms in children. *Personality and Individual Differences*, 54(3), 367-371.
- Van Doorn, R. R., & Lang, J. W. (2010). Performance differences explained by the neuroticism facets withdrawal and volatility, variations in task demand, and effort allocation. *Journal of Research in Personality*, 44(4), 446-452.
- Vasilopoulos, N. L., Cucina, J. M., & McElreath, J. M. (2005). Do warnings of response verification moderate the relationship between personality and cognitive ability? *Journal of Applied Psychology*, 90(2), 306-322.

- Vassend, O., & Skrandal, A. (2011). The NEO personality inventory revised (NEO-PI-R): Exploring the measurement structure and variants of the five-factor model. *Personality and Individual Differences*, 50(8), 1300-1304.
- Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, 59(2), 197-210.
- Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, 83(2), 213-217.
- Walmsley, P. T., & Sackett, P. R. (2013). Factors affecting potential personality retest improvement after initial failure. *Human Performance*, 26(5), 390-408.
- Weijters, B., Geuens, M., & Schillewaert, N. (2009). The proximity effect: The role of inter-item distance on reverse-item bias. *International Journal of Research in Marketing*, 26(1), 2-12.
- White, L. A., Young, M. C., Hunter, A. E., & Rumsey, M. G. (2008). Lessons learned in transitioning personality measures from research to operational settings. *Industrial and Organizational Psychology*, 1(03), 291-295.
- Wilk, S. L., Desmarais, L. B., & Sackett, P. R. (1995). Gravitation to jobs commensurate with ability: Longitudinal and cross-sectional tests. *Journal of Applied Psychology*, 80(1), 79-85.
- Williams, L. J., Hartman, N., & Cavazotte, F. (2010). Method variance and marker variables: A review and comprehensive CFA marker technique. *Organizational Research Methods*, 13(3), 477-514.
- Zaccaro, S. J., Gulick, L. M., & Khare, V. P. (2008). Personality and leadership. *Leadership at the Crossroads*, 13-29.

- Zawadzki, B., & Strelau, J. (2010). Structure of personality: Search for a general factor viewed from a temperament perspective. *Personality and Individual Differences*, 49(2), 77-82.
- Zhang, T., Gino, F., & Bazerman, M. H. (2014). Morality rebooted: Exploring simple fixes to our moral bugs. *Research in Organizational Behavior*, 34, 63-79.
- Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods*, 7(2), 168-190.
- Ziegler, M., & Buehner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement*, 69 (4), 548-565
- Ziegler, M., MacCann, C., & Roberts, R. D. (2012). Faking: Knowns, unknowns, and points of contention. *New Perspectives on Faking in Personality Assessment*, 3-16.
- Ziegler, M., Schmidt-Atzert, L., Bühner, M., & Krumm, S. (2007). Fakability of different measurement methods for achievement motivation: Questionnaire, semi projective, and objective. *Psychology Science*, 49(4), 291–307.