

**Ludwig-Maximilians-Universität**

**Lehrstuhl für Psychologische Methodenlehre und Evaluation**

**Professor Dr. Markus Bühner**



**Situational Demand and Its Impact on Construct and Criterion**

**Validity of a Personality Questionnaire:**

**State and Trait, a Couple you just can't Study Separately!**

Matthias Ziegler

München 2006

**Betreuer der Dissertation:**

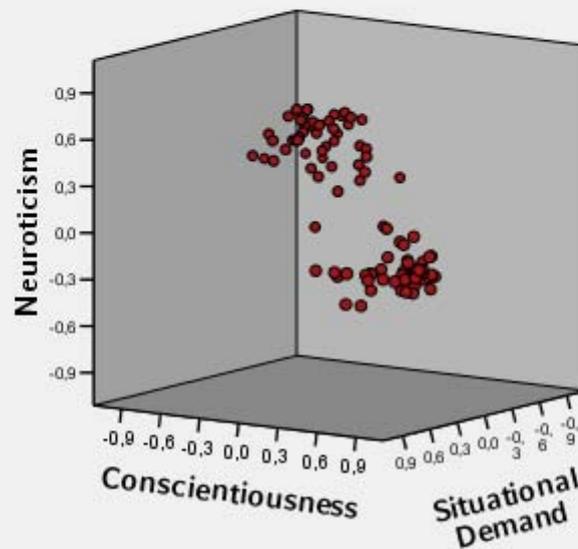
Referent: Prof. Dr. Markus Bühner  
Koreferent: Prof. Dr. Reinhard Pekrun  
Tag der mündlichen Prüfung: 24.01.2007

© Matthias Ziegler, 2006  
Alle Rechte - vor allem die zur Vervielfältigung und Verbreitung  
(auch über elektronische Medien) - liegen beim Autor.  
Dies gilt natürlich nicht für Elemente, die aus anderen Veröffentlichungen  
entnommen sind.

## Situational Demand and Its Impact on Construct and Criterion

### Validity of a Personality Questionnaire:

### State and Trait, a Couple you just can't Study Separately!



Inaugural – Dissertation  
zur Erlangung des Doktorgrades  
der Philosophie an der Ludwig – Maximilians – Universität  
München

vorgelegt von

Matthias Ziegler

September 2006

*“... we always measure persons in situations, not persons; there is no psychological measurement in the situational vacuum.”*

Deinzer et al. (1995)

## Acknowledgments

I want to use this space to express my gratitude for a number of people without whom this dissertation could not have been finished.

First of all I want to thank Prof. Dr. Markus Bühner. Markus, I never regretted that I went to Munich with you. You always supported and inspired me. There are so many things I have learned from you and still can learn. Therefore, I hope, our quest will continue.

I also want to thank Prof. Dr. Friedrich Försterling who helped me during the beginnings of this project. Moreover, I am indebted to Prof. Dr. Reinhard Pekrun and Prof. Dr. Hartmut Ditton for their willingness to function as jury in this trial.

Furthermore, I want to thank Prof. Dr. Rolf Steyer. Even though it was not his intention, he gave me the right advice at the right time.

Of course, I want to thank my family, especially my parents. I will never be able to repay you properly. I hope I will make you proud.

I do not want to forget to thank the students who helped to realize this project: Anne – Kathrin Gerber, Natalia Zorkina, and Yuri Bassiakov have to be mentioned, but all participants deserve my gratitude as well.

Annika Mörsdorf was kind enough to proof read the first draft and helped to improve it significantly.

Last but not least, I want to thank Daniela Otto. You were there and helped whenever I was down or felt restive. It is my wish that this will always be so!

Matthias Ziegler

## Zusammenfassung

Persönlichkeit ist ein Schlagwort in der heutigen Gesellschaft, das häufig und in sehr vielen Zusammenhängen eingesetzt wird. Auch in der psychologischen Forschung spielt die Persönlichkeit eine große Rolle. Es existieren zahlreiche Persönlichkeitstheorien und –modelle. Diese überschneiden sich zum Teil sehr stark. Allerdings gibt es auch konkurrierende Forschungsansätze. So können der typologische und der dimensionale Ansatz gegeneinander abgegrenzt werden.

Typologien der Persönlichkeit bestehen seit der Antike. Hier wurden vier Persönlichkeitstypen gemäß der vorherrschenden Körperflüssigkeit unterschieden. Im Gegensatz zu Typologien wird bei dimensional Ansätzen von einer Reihe unabhängiger Persönlichkeitseigenschaften ausgegangen. Unabhängig bezieht sich darauf, dass die individuelle Ausprägung auf einer der Dimensionen keinen Einfluss auf die individuelle Ausprägung auf einer der anderen Dimensionen hat. Rein theoretisch ist es also möglich, mit einem dimensional Ansatz unendliche viele Persönlichkeitsprofile zu erstellen.

Lange Zeit herrschte eine große Debatte darüber, wie viele verschiedene Dimensionen sinnvoll seien, um die Persönlichkeit eines Menschen zu beschreiben. Mit Hilfe von Selbstbeschreibungen und faktorenanalytischen Methoden wurde versucht, die Anzahl systematisch einzugrenzen. Nach einem langen Diskurs, dessen Ende immer noch nicht erreicht ist, scheint eine Lösung jedoch immer wieder aufzutauchen. Hierbei handelt es sich um die sogenannten Big 5: Neurotizismus, Extraversion, Offenheit für Erfahrung, Verträglichkeit und Gewissenhaftigkeit.

Ein kurzer Abriss über die Entstehungsgeschichte der Big 5 ist in der Einleitung dieser Arbeit gegeben.

Die prominentesten Befürworter der Big 5 sind Costa und McCrae (1992b). Ihr Fragebogen, das NEO – PI – R gehört zu den am häufigsten eingesetzten Verfahren in Wissenschaft und Praxis.

Jedes psychometrische Verfahren, welches als anerkannt und sinnvoll gelten will, muss verschiedene Gütekriterien erfüllen. Zunächst muss die Erfassung objektiv und reliabel sein. Da es sich bei den meisten Instrumenten zur Erfassung der Big 5 um standardisierte Fragebogenverfahren handelt, kann die Objektivität als gesichert gelten. Auch für die Reliabilität liegen zahlreiche positive Ergebnisse vor. Bei der Validität können drei verschiedene Aspekte unterschieden werden: Inhalts-, Konstrukt- und Kriteriumsvalidität. Die Inhaltsvalidität eines Verfahrens lässt sich nicht statistisch absichern, sondern nur theoretisch bestimmen. Ein Verfahren ist dann inhaltsvalide, wenn es aus dem theoretisch möglichen Itempool eine repräsentative Stichprobe an Items enthält. Mit anderen Worten heißt das, es soll das zu messende Konstrukt möglichst komplett erfassen. Die immer noch schwelende Diskussion um die Anzahl der Faktoren ist sicher teil der Bestimmung der Inhaltsvalidität. Die vorliegende Arbeit war jedoch nicht mit dieser Thematik beschäftigt. Vielmehr wurden Fragen der Konstrukt- und der Kriteriumsvalidität untersucht.

Wie bereits erwähnt, wird in der Theorie angenommen, dass die Persönlichkeitsdimensionen unabhängig voneinander sind. Im statistischen Sinne würde das bedeuten, dass die Verfahren nicht miteinander korrelieren dürfen. In einer Vielzahl von Studien zeigten sich jedoch immer wieder substantielle Zusammenhänge zwischen Persönlichkeitsdimensionen. In einer Metaanalyse von 1995 fanden Mount, Barrick, Scullen und Rounds bereinigte Zusammenhänge bis zu .52 zwischen Neurotizismus und Gewissenhaftigkeit. Costa und McCrae (1992b) verschweigen diese Zusammenhänge ebenfalls nicht, führen sie jedoch auf fehlende

diskriminante Validität einiger Facetten zurück. Weiterhin empfehlen Sie, Faktorwerte zu berechnen anhand der Faktorgewichte, die sich in der amerikanischen Normstichprobe ergeben haben. Hier zeigte sich jedoch eine bivariate Korrelation von .53 zwischen Neurotizismus und Gewissenhaftigkeit. Diese ist also noch größer als die bereinigte Korrelation, die von Mount und Kollegen berichtet wurde. Es stellt sich also die Frage, ob die Big 5 tatsächlich unkorreliert sind oder doch substantielle Zusammenhänge bestehen. Diese Frage wurde in der ersten und in der zweiten Studie der vorliegenden Arbeit untersucht.

Innerhalb beider Studien wurde ein neuer Erklärungsansatz für die Zusammenhänge untersucht. Die Idee dabei ist, dass die Zusammenhänge durch einen situativen Einfluss entstehen. Dieser situative Einfluss (situational demand) setzt sich im Sinne eines korrelierten Fehlers auf die Varianz der Persönlichkeitswerte. Da der situative Einfluss alle Variablen gleich betrifft, steigen die Korrelationen zwischen den Faktoren proportional zum situativen Druck. In einem Bewerbersetting müsste der situative Druck beispielsweise deutlich höher sein als in einer anonymen Laborsituation.

Mit Hilfe der Latent – State – Trait – Theorie (Steyer, Ferring & Schmitt, 1992) und den individuellen kausalen Effekt Design (Steyer, 2005) wurde untersucht, welchen Einfluss die Situation auf die Korrelationen zwischen den Persönlichkeitsfaktoren hat.

Bei der Kriteriumsvalidität handelt es sich um die Frage, ob die Big 5 in der Lage sind, relevante Kriterien vorherzusagen. So wurde der Zusammenhang zu Berufserfolg (z.B. Schmidt & Hunter, 1998; Salgado, 2003) und akademischen Erfolg untersucht (z.B. Furnham, Chamorro Premuzic & McDougall, 2002). Die gefundenen Zusammenhänge sind zwar zumeist gering, allerdings zeigen sie sich stabil und inkrementell zu anderen Konstrukten wie Intelligenz.

Trotz dieser Argumente werden Persönlichkeitsfragebogen von einigen Praktikern nur zögerlich oder ungern eingesetzt. Grund dafür sind Bedenken über die Verfälschbarkeit dieser Instrumente. Die entscheidende Frage ist jedoch, woher kommt die prädiktive Kraft eines Fragebogens, wenn er verfälscht wurde? Diese Frage wurde ebenfalls in Studie 2 untersucht. Mit Hilfe des beschriebenen Designs ist es möglich, die Varianz aufzuspalten in Varianz, die auf Persönlichkeit zurückgeht und Varianz, die auf Verfälschung zurückgeht. Die prädiktiven Anteile können dann getrennt untersucht werden.

Die meisten Untersuchungen, die sich mit Verfälschung (Faking) beschäftigen, gehen davon aus, dass Faking ein uniformer Vorgang ist, der von allen Personen gleich ausgeführt wird (z.B. Pauls & Crost, 2005b). Allerdings zeigen neuere Untersuchungen, die Techniken aus der Item – Response – Theorie nutzen, dass verschiedene Fakingstile unterschieden werden können (Zickar & Robie, 1999; Zickar, Gibby & Robie, 2004). So gibt es trotz der Aufforderung zum Faken immer noch eine geringe Anzahl von Leuten, die nicht lügen, sondern ehrlich antworten (regular responders). Eine weitere Gruppe lügt extrem (extreme faker). Der größere Anteil der Personen verzerrt ihre Antwort jedoch nur gering (slight faker). Diese Tatsache legt die Frage nahe, ob sich die einzelnen Stile in ihrer Kriteriumsvalidität unterscheiden. Diese Frage wurde in der 3. Studie untersucht.

Eine weitere Frage, die sich stellt, ist die Frage nach dem psychologischen Prozess, der stattfindet, wenn Personen ihre Antworten absichtlich verfälschen. Auch diese Frage wurde in der 3. Studie mit Hilfe einer qualitativen Untersuchung näher beleuchtet.

Die Ergebnisse der drei Studien werden im Folgenden kurz dargestellt.

*Studie 1.* In der ersten Studie wurden Daten reanalysiert, die Toomela bereits 2003 publiziert hat. In dieser Stichprobe wurden  $N = 912$  Männer des estnischen Militärs mit der estnischen Version des NEO – PI – R und einem Intelligenztest untersucht. In seiner Originalarbeit berichtete Toomela, dass in der Gruppe der intelligentesten Männer, im Gegensatz zu den anderen Intelligenzgruppen, eine vierfaktorielle Lösung besser paste als ein fünffaktorielle Lösung. Der Grund hierfür war eine starke Korrelation zwischen Neurotizismus und Gewissenhaftigkeit. Dadurch konnten diese beiden Dimensionen nicht voneinander getrennt werden.

In der hier durchgeführten Reanalyse wurde untersucht, ob sich dieses Problem auf einen situativen Einfluss zurückführen lässt. Im Laufe der Reanalyse zeigte sich, dass nicht Intelligenzunterschiede, sondern vielmehr Unterschiede im militärischen Rang eine Rolle bei der Entstehung der unterschiedlichen Faktorenlösungen spielte. Tatsächlich bestand die Stichprobe aus Rekruten, Unteroffizieren und Offizieren. Die Annahme, die hier getroffen wurde, lautete, dass der situative Druck, den die Probanden verspürten mit dem militärischen Rang anstieg.

Um dies zu testen, und gleichzeitig zu prüfen, ob die starke Korrelation sinkt, wenn der situative Einfluss kontrolliert wird, wurde ein Multigruppen – Strukturgleichungsmodell berechnet. Dieses Modell enthielt die beiden latenten Persönlichkeitsfaktoren Neurotizismus und Gewissenhaftigkeit sowie einen weiteren latenten Faktor (fake), der den situativen Einfluss repräsentiert. Während die beiden Persönlichkeitsfaktoren nur von ihren jeweiligen Facetten Ladungen erhalten, zieht der situative Faktor Varianz von allen Facetten. Dies ist die Umsetzung der Idee, dass situativer Einfluss wie ein systematischer, korrelierter Fehler wirkt.

Die Ergebnisse der Reanalyse zeigen, dass die Korrelation tatsächlich stark zurückgeht, sich sogar umdreht, wenn situativer Einfluss kontrolliert wird. Somit liegt eine erste Evidenz vor, dass ein unkorreliertes Modell der Big 5 nicht

unrealistisch ist, sofern der situative Einfluss kontrolliert wird. Darüber hinaus zeigte sich erwartungskonform, dass die Offiziere den höchsten Mittelwert in der latenten Variable Fake erzielten. Sie verspürten also den größten situativen Druck.

Einschränkend muss festgehalten werden, dass die hier durchgeführte Analyse lediglich aus einem Messzeitpunkt bestand. Dadurch ist eine reine Trennung von Persönlichkeits- und Situationsvarianz nicht möglich. Um diese Trennung zu erzielen, wurde die zweite Studie durchgeführt.

*Studie 2.* In dieser Studie wurden  $N = 186$  zufällig auf zwei Gruppen verteilt. Probanden in der ersten Gruppe (KG) bearbeiteten das NEO – PI – R zweimal, mit der Aufforderung, ehrlich zu antworten. In der zweiten Gruppe (EG) wurde das NEO – PI – R beim ersten Mal ehrlich und beim zweiten Mal mit einer Verfälschungsinstruktion durchgeführt. Zusätzlich füllten beide Gruppen weitere Fragebogen sowie einen Intelligenzstrukturtest aus. Durch dieses Design war es möglich, die Varianz in Persönlichkeits- und Situationsvarianz zu spalten. Darüber hinaus konnte der Einfluss des situativen Drucks auf die Korrelationen zwischen den Persönlichkeitsdimensionen untersucht werden. Da alle Probanden auch an einer später durchgeführten Statistikklausur teilnahmen, konnte ebenfalls die Kriteriumsvalidität der beiden Anteile bestimmt werden.

Die Ergebnisse der Untersuchung zeigen, dass die Korrelationen zwischen den Faktoren tatsächlich deutlich sinken, wenn der situative Druck kontrolliert wird. Lediglich Extraversion und Verträglichkeit wiesen noch einen substanziellen Zusammenhang auf. Demzufolge kann geschlussfolgert werden, dass der situative Einfluss für einen großen Teil der Zusammenhänge zwischen Dimensionen verantwortlich ist. Dennoch zeigte sich bei einigen Facetten eine mangelnde diskriminante Validität.

Bezüglich des Charakters der Fakingvariable zeigte sich der stärkste Zusammenhang mit Selbstwirksamkeitsüberzeugungen. Personen, die glauben, andere so beeindruckend zu können, dass sie als kompetent und liebenswürdig wahrgenommen werden, verfälschten auch am meisten. Allerdings war der Zusammenhang gering, so dass auch andere Variablen bei der Verfälschung eine Rolle spielen müssen.

Schließlich konnte gezeigt werden, dass die prädiktive Kraft, selbst unter Fakingbedingungen, immer noch von der Persönlichkeitsvariable kommt. Jedoch hat diese Variable nur noch Ladungen von ein oder zwei Facetten. Zusammengefasst heißt das, dass die prädiktive Kraft nicht mehr von einer Dimension, sondern von den Facetten stammt. Faking hatte nur eine sehr geringe, wenn auch positive prädiktive Kraft.

Aus diesen Ergebnissen wurde geschlossen, dass spezifisches Wissen über die Anforderungen eine Rolle spielen muss, wenn Personen Fragebogen verfälschen.

Eine Einschränkung dieser Studie war, dass Faken als uniformer Prozess aufgefasst wurde. Deshalb wurde in Studie 3 der psychologische Prozess, der beim Verfälschen abläuft, näher beleuchtet.

*Studie 3.* Im ersten Teil der Studie wurde mit Hilfe von kognitiven Interviews und einer Stichprobe von  $N = 50$  untersucht, welche Strategien Personen anwenden, wenn sie ihre Antworten in einem Fragebogen absichtlich verfälschen. Die Auswertung, die durch zwei Rater durchgeführt wurde, zeigte, dass zwei Verfälschungsstrategien unterschieden werden können: leichte Verfälschung und extreme Verfälschung. Ein weiteres, interessantes Ergebnis war, dass nicht alle Items verfälscht wurden, sondern nur diejenigen, die von den Probanden als wichtig

angesehen wurden. Für diese Einstufung wurden implizite Theorien aber auch spezifisches Wissen genutzt. Bei den Fragen, die nicht als wichtig eingestuft wurden, antworteten die Probanden entweder ehrlich oder neutral.

In einem weiteren Teil der Untersuchung wurde die Stichprobe aus der zweiten Studie mit der neuen Stichprobe kombiniert und mit Hilfe von Mixed Rasch Modellen wurde nach der Anzahl von unterschiedlichen Fakingklassen gesucht. Hier zeigten sich drei Klassen: extreme faker, slight faker und regular responders.

Weitere Analysen zeigten, dass die leichten Verfälscher über eine geringere kognitive Leistungsfähigkeit und höheren Neurotizismus verfügen. In beiden Verfälschungsgruppen konnte immer noch eine starke Basis an ehrlicher Persönlichkeit nachgewiesen werden. Interessanterweise hatten extreme Faker höhere Werte im Schlußfolgernden Denken und in der Klausur. Als Schlussfolgerung wurde gezogen, dass Personen, die nur leicht verfälschen, nicht nur Angst haben, entdeckt zu werden, sondern auch geringe Fähigkeiten haben, zu verfälschen.

Weiterhin konnte gezeigt werden, dass die Kriteriumsvalidität in allen drei Klassen vergleichbar ist. Allerdings fungierte Klassenzugehörigkeit als Prädiktor für die Klausurleistung, auch, wenn Intelligenz kontrolliert wurde.

Abschließend wurden die verschiedenen Ergebnisse der einzelnen Studien integriert und ein Modell zur Erklärung des Einflusses von situativem Druck auf die Beantwortung von Fragebogen vorgeschlagen.

Das Modell besteht aus zwei Hauptpfaden, die für hohen und niedrigen situativen Druck stehen. Weiterhin wird dargestellt, wie es zu den unterschiedlichen Antwortstilen kommt und wie diese sich auf die Konstrukt- und die Kriteriumsvalidität eines Fragebogens auswirken.

---

<b>CHAPTER 1. INTRODUCTION</b>	<b>5</b>
1.1. A SHORT HISTORY OF THE BIG 5	6
1.2. APPLICATIONS OF THE BIG 5	9
1.2.1. <i>Use in Organizational Psychology</i>	10
1.2.1.1. Predicting job performance using the BIG 5	11
1.2.1.2. Predicting academic success using the BIG 5	12
1.2.2. <i>Use in Clinical Psychology</i>	13
1.2.2.1. Big 5 and Alexithymia	13
1.2.2.2. Big 5 and Depression	14
1.2.2.3. Big 5 and Eating Disorder	15
1.2.2.4. Big 5 and Schizophrenia	15
1.2.2.5. Big 5 and Personality Disorders	16
1.2.3. <i>Use in Pedagogic Psychology</i>	17
1.2.4. <i>Summary</i>	18
1.3. CONCERNS ABOUT THE BIG 5	18
1.3.1. <i>Number of Factors</i>	18
1.3.2. <i>Methodological Issues</i>	20
1.3.3. <i>Orthogonality</i>	21
1.3.4. <i>Trait and State Variance</i>	22
1.3.5. <i>Social Desirability and Its Impact on Construct and Criterion Validity</i>	24
1.3.6. <i>Summary</i>	26
1.4. GOALS OF THE PRESENT PROJECT	28
1.4.1. <i>Impact of Situational Demand on the Construct Validity of the BIG 5</i>	28
1.4.2. <i>Impact of Situational Demand on the Criterion Validity of the BIG 5</i>	31
1.4.3. <i>The Psychological Process behind Situational Demand</i>	33
1.4.4. <i>Dealing with Situational Demand</i>	34
1.4.5. <i>Summary and outlook</i>	36
<b>CHAPTER 2. A REANALYSIS OF TOOMELA (2003): PERSONALITY STRUCTURE - A QUESTION OF IQ OR SITUATIONAL DEMAND?</b>	<b>38</b>
2.1. THE STUDY BY TOOMELA	43
2.2. REANALYSIS	44
2.3. STATISTICAL ANALYSES	49
2.4. RESULTS	50
2.4.1. <i>Sample description</i>	50
2.4.2. <i>Structural equation analysis</i>	51
2.5. DISCUSSION	57
2.5.1. <i>Situational demand</i>	58
2.5.2. <i>BIG 5 Models</i>	60
2.5.3. <i>Relationship between personality dimensions</i>	61
2.5.4. <i>Latent personality means</i>	62

---

2.5.5.	<i>Limitations and further research</i>	62
<b>CHAPTER 3. PERSONALITY STRUCTURE: A QUESTION OF TRAIT &amp; STATE INTERACTION</b>		<b>66</b>
3.1.	CRITIC ON THE CONSTRUCT VALIDITY OF THE FFM	66
3.2.	CRITIC ON THE CRITERION VALIDITY OF THE BIG 5	69
3.3.	SUBSTANCE VS. STYLE	71
3.4.	PRESENT STUDY	72
3.5.	METHODS	74
3.5.1.	<i>Participants and procedure</i>	74
3.5.2.	<i>Test materials</i>	75
3.5.3.	<i>Models</i>	77
3.5.4.	<i>Statistical Analysis</i>	78
3.6.	RESULTS	81
3.6.1.	<i>Without controlling situational demand</i>	81
3.6.2.	<i>Controlling situational demand</i>	85
3.7.	DISCUSSION	90
3.7.1.	<i>Construct validity</i>	91
3.7.2.	<i>Substance vs. style</i>	93
3.7.3.	<i>Criterion validity</i>	95
3.7.4.	<i>A hypothesis regarding faking</i>	95
<b>CHAPTER 4. SITUATIONAL DEMAND, RESPONSE SETS, AND FAKING IN THE LIGHT OF COGNITIVE INTERVIEWS AND MIXED RASCH MODELS</b>		<b>99</b>
4.1.	THE PRESENT STUDY	102
4.1.1.	<i>Psychological processes</i>	102
4.1.2.	<i>Replication of response classes</i>	103
4.1.3.	<i>Causes for differences in responding to situational demand</i>	104
4.1.4.	<i>Criterion validity</i>	104
4.2.	METHOD	106
4.2.1.	<i>Sample and procedure</i>	106
4.2.2.	<i>Test materials</i>	108
4.2.3.	<i>Statistical analysis</i>	109
4.3.	QUALITATIVE ANALYSIS	112
4.3.1.	<i>Results</i>	112
4.3.2.	<i>Discussion</i>	113
4.4.	QUANTITATIVE ANALYSIS	116
4.4.1.	<i>Replication of faking classes</i>	116
4.4.2.	<i>Causes for differences in responding to situational demand</i>	120
4.4.3.	<i>Criterion validity</i>	121
4.5.1.	<i>Response classes</i>	126
4.5.2.	<i>Model of responding to situational demand</i>	128
4.5.3.	<i>Practical implications</i>	131

---

4.5.4.	<i>Methodological implications</i>	133
4.6.	LIMITATIONS	133
<b>CHAPTER 5_ CONCLUSIONS</b>		<b>136</b>
5.1.	SUMMARIES OF THE THREE STUDIES PRESENTED	136
5.1.1.	<i>Study 1</i>	136
5.1.2.	<i>Study 2</i>	138
5.1.3.	<i>Study 3</i>	140
5.2.	CONCLUSIONS REGARDING THE GOALS OF THE PRESENT PROJECT	143
5.2.1.	<i>Impact of Situational Demand on the Construct Validity of the BIG 5</i>	143
5.2.2.	<i>Impact of Situational Demand on the Criterion Validity of the BIG 5</i>	145
5.2.3.	<i>The Psychological Process behind Situational Demand</i>	147
5.2.4.	<i>Dealing with Situational Demand</i>	149
5.3.	SUMMARY AND OUTLOOK	150
<b>REFERENCES</b>		<b>152</b>

## **Chapter 1: Introduction**

## Chapter 1. Introduction

Why are some people more courageous than others? Why did a person choose to become a leader? And why was that person a stronger and more efficient leader than somebody else? Scientists and laypeople alike have searched for answers to these questions for a long time now. The reasons for this interest are the need to understand, predict, and alter human behavior. In ancient Greece actors in theater plays wore masks to demonstrate the specific character of the role they played (Allport, 1961). In order to do this, the first studies of typical characteristics of a person were undertaken. Moreover, Hippocrates hypothesized that there are four distinct body fluids, namely blood, phlegm, black bile, and yellow bile which determine the character of a person. This description can be regarded as the first personality typology. Typologies of personality still enjoy popularity today (e.g., Furnham, Moutafi & Paltiel, 2005) but have been broadly criticized lately (Asendorpf, 2006; McCrae, Terracciano, Costa & Ozer, 2006a; McCrae, Terracciano, Costa & Ozer, 2006b). The main criticism has been that reported personality types are not replicable, however, personality dimensions are.

Personality dimensions, also called domains, have become the most popular construct for the description of personality. According to Liebert and Liebert (1998, p. 5-6) personality as a whole can be defined as: "... the unique, dynamic organization of characteristics of a particular person, physical and psychological, which influence behavior and responses to the social and physical environment. Of these characteristics, some will be entirely unique to the specific person (i.e. memories, habits, mannerisms) and others will be shared with a few, many, or all other people". A dimensional view of personality mostly uses the trait concept to

describe common differences between individuals. Traits can be seen as “Enduring characteristics on which individuals differ” (Liebert & Liebert, 1998, p. 184).

In 1936 Allport and Odbert published a list of 17.953 single – word descriptors they had found in a dictionary and which could be used to describe a person (Allport & Odbert, 1936). Cattell (1943a) took up this idea and formulated what was to become known as the lexical hypothesis: "All aspects of human personality which are or have been of importance, interest, or utility have already become recorded in the substance of language" (1943b p. 483). The lexical hypothesis has subsequently been the starting point for several different personality theories. The most prominent of these theories is the Five Factor Model (FFM) which is also called the BIG 5.

The BIG 5 have been chosen as research object within the present analyses because they rank among the most prominent psychological constructs in research and practice. The present paper will give an introduction into the different applications of the BIG 5 in chapter 1.2. However, the BIG 5 have also been criticized on a number of issues. These issues will shortly be presented in chapter 1.3. Finally, the goals of the present project which aim at answering some of the criticisms regarding the BIG 5 are presented in chapter 1.4. As a beginning, though, a short overview of the history of the BIG 5 is provided.

### 1.1. A Short History of the BIG 5

As mentioned above the pioneer work by Allport and Odbert (1936) has been the starting point for a branch of personality research which is also known as factor analytical personality research. Factor analysis is a powerful methodological instrument which was developed for data reduction and moreover, to find latent

factors responsible for correlations between variables. These latent factors also were the object of interest in the first factor analytical studies of personality by Cattell (1943a, b). Cattell formulated the lexical hypothesis. He also acknowledged that the original list of words by Allport and Odbert was insufficient. Thus, he added words which he believed to be important descriptors of personality and which also represented the state of the art of the research of that time. Using personal judgement as well as correlational techniques Cattell limited the vast number of words to a list of 35 bipolar adjectives. This list was administered as peer evaluation and the resulting data set was factor analyzed. Cattell reasoned that applying factor analysis to the data set would help to find the latent personality dimensions which are used to describe a person. His analysis came up with twelve personality factors. In the forthcoming years this adjective list has been used in many different research projects, even though Cattell himself acknowledged the fact that he might have eliminated important words.

However, this adjective list was also used by Tupes and Christal who were the first to report five personality factors. Tupes and Christal worked for the US Air Force. Their task was to improve officer selection and promotion procedures. In eight different factor analyses they always found five replicable factors even though the total number of factors in each analysis differed. Later these studies were regarded as the discovery of the BIG 5 (McCrae, 1992). Since Tupes and Christal did not publish their findings until 1992 they would not have impacted the scientific community had not Norman (1963) taken up the idea of five factors. However, Norman (1967) also decided to go back to the total variety of words and made a new list of 1.431 words (for a more detailed discussion see John, Angleitner & Ostendorf, 1988). He ordered them into 75 semantic clusters which could be attributed to one

of the five factors. By then the first wave of factor analytical studies of personality came to an end.

The second research wave began with Goldberg (e.g., 1981; Goldberg, 1992) and still runs until today. Goldberg again started using more adjectives and repeatedly found five factors. He coined the name BIG 5 which today stands for the five personality factors deemed to be necessary to describe the covariances in individual self ratings. Yet, the most prominent proponents of the BIG 5 are Costa and McCrae. They used the adjective lists and formulated concrete items. Originally, Costa and McCrae investigated the BIG 2 (Neuroticism and Extraversion) and later the BIG 3 (Openness was added). However, inclined to develop an instrument to assess five personality factors they broadened their item pool to incorporate Agreeableness and Conscientiousness into their model. Furthermore, Costa and McCrae were not satisfied with only five relatively broad personality factors. They reasoned that these factors must comprise lower order personality aspects which they called facets. Accordingly, they came up with six facets for each of the five factors. At this point, it must be mentioned that Costa and McCrae left the path of pure factor analytical research. The facets were based on the mutual understanding of both scientist and thus, merely represent their views. The facets were each assessed with eight items. Thus, the final questionnaire consisted of 240 items assessing the BIG 5 and their corresponding facets (Costa & McCrae, 1992b). Even though the facet structure was not derived from factor analyses the following research proved that the proposed model fitted the data well. Therefore, Goldberg (1992) concluded: "Gradually, agreement has been growing about the number of orthogonal factors needed to account for the interrelations among English-language trait descriptors" (p. 26). Those factors are the BIG 5, namely, Neuroticism, Extraversion, Openness, Agreeableness, and Conscientiousness. According to Costa

and McCrae Neuroticism comprises such facets as anxiety, depression, impulsiveness, and vulnerability. Therefore, it has also been labelled Emotional Stability. Extraversion contains individual differences regarding activity, gregariousness, positive emotions, and excitement seeking. Openness to fantasy, ideas, values, and feelings are some of the facets which are a part of Openness (to Experience). Agreeableness includes facets like trust, altruism, compliance, and modesty. Finally, Conscientiousness comprises amongst others competence, order, dutifulness, and achievement striving<sup>1</sup>.

Several critics have raised concerns regarding the BIG 5. Some of these concerns will be discussed later. Yet, the evidence supporting the BIG 5 model seems overwhelming. The model has been replicated in different cultures and using different questionnaires (e.g., Digman & Shmelyov, 1996). Moreover, studies have shown that estimates of heritability range from .39 for agreeableness to .49 for extraversion (Bouchard, 1997). Besides this, there is evidence for the stability of BIG 5 scores (e.g., Costa & McCrae, 2006; Roberts, Walton & Viechtbauer, 2006). The BIG 5 model has proven its value in numerous other studies and fields. A short overview of these different applications will be given next.

## 1.2. Applications of the BIG 5

As soon as the majority of the scientific community had agreed upon the BIG 5 as the most promising personality model it was tested in several fields. Overall, virtually no psychological field of research seems to remain where the BIG 5 have not been applied. The range covers personality of web page owners (Marcus, Machilek & Schutz, 2006), interview competencies of police officers (De Fruyt,

---

<sup>1</sup> From here on the personality factors will be written without a capital letter at the beginning to improve legibility.

Bockstaele, Taris & Van Hiel, in press), and connections to physiological research (Smits & Boeck, 2006). To address all the different fields of research in which the BIG 5 have been used would clearly be beyond the scope of the present paper. Therefore, examples for the value of the BIG 5 within three selected areas of psychology will be given next.

### *1.2.1. Use in Organizational Psychology*

Within organizational psychology personality assessment serves several purposes. In personnel development skills of employees are supposed to be trained and increased. Here personality assessment can help to select those employees who bring the necessary motivation and openness along. Looking at teamwork, there always seems to be the question of how to mix a team in terms of personalities (Peeters, Rutte, van Tuijl & Reymen, 2006). Recent meta analytical analyses show that agreeableness can be regarded as the most important personality predictor regarding the outcome of teams. Psychologists have also tried to find out for a long time which traits enable a person to become a successful leader (e.g., Bono & Judge, 2004; e.g., Shao & Webber, 2006). Bono and Judge found extraversion was the strongest and most consistent correlate of transformational leadership. However, overall correlations between traits and leadership were rather small.

Yet, one of the most common fields of applications and the field of most interest for the present paper is personnel selection.

### 1.2.1.1. Predicting job performance using the BIG 5

Pioneers in the field were Barrick and Mount (1991). They realized that the various and partly contradicting results for the predictive power of personality within personnel selection might in part be due to the confusing variability in names and constructs used to assess personality. At about the same time meta – analyses became very popular due to the works by Schmidt and Hunter (Hunter & Hunter, 1984; Hunter & Schmidt, 1990; Schmidt & Hunter, 1998). Barrick and Mount recognized their chance and used the BIG 5 model to order the different constructs applied and moreover, they applied meta – analytical techniques to combine the findings. Their article published in 1991 was the most cited psychological article in the 1990ies (Mount & Barrick, 1998). The results revealed conscientiousness as the only personality factor which showed consistent relations with three different job performance criteria (job proficiency, training proficiency, and personnel data) over five different occupational groups (professionals, police, managers, sales, and skilled/semi-skilled). Extraversion also proved to be a valid predictor for training proficiency across the different occupations. Even though most validity coefficients were below .20 the results demonstrated that personality is a valid predictor of performance within occupational settings. These findings inspired many other researchers to further the knowledge base.

Salgado (1997) was the first who used meta – analyses to combine research evidence collected in Europe. His results were identical to Barrick and Mount's with the exception that he also found neuroticism to be a valid predictor across different jobs and different criteria. The actual validities reported were again rather small, only conscientiousness achieved a higher true correlation with training success of .39. Later Salgado (2003) demonstrated that BIG 5 – based questionnaires generally

reach higher criterion validities than non BIG 5 – based questionnaires. This lends further support to the claim that the BIG 5 are valid predictors of performance and preferable to other personality models.

Judge and Ilies (2002) were able to show that neuroticism (average validity = -.31) and conscientiousness (average validity = .24) predict performance motivation (goal-setting, expectancy, and self-efficacy motivation). Thus, the same two traits of all five traits which Salgado found attributed the most to job performance also explain motivational aspects within occupational settings. Moreover, Judge and Ilies computed that the BIG 5 as a whole explain about 24 percent of the variance in performance motivation. This finding can be regarded as exceptional.

#### 1.2.1.2. Predicting academic success using the BIG 5

Beyond the prediction of job performance, the BIG 5 have also been used to predict performance in academic settings (e.g., Chamorro Premuzic & Furnham, 2003; Farsides & Woodfield, 2003; Furnham & Chamorro-Premuzic, 2004; e.g., Furnham, Chamorro Premuzic & McDougall, 2002; Rindermann & Neubauer, 2001). Furnham and colleagues investigated the incremental validity of personality above and beyond intelligence. In 2004 they reported that personality explained twelve percent of the variance in a statistic examination grade above and beyond intelligence within a sample of university students. Extraversion and conscientiousness were the most important personality predictors. Interestingly, higher extraversion was counterproductive for a good result in the statistic examination. In a longitudinal study over three years Farsides and Woodfield found that the BIG 5 explained about six percent of the variance in the final undergraduate

grade above and beyond intellect and motivation. However, only openness reached a significant regression weight ( $\beta = .25$ ).

Thus, the evidence concerning the criterion validity of the BIG 5 in predicting academic success is positive. However, it remains unclear which traits are necessary in which field.

Summing up, the conclusion is that the BIG 5 are valid predictors for performance in various fields. Certainly, they are not the strongest predictors because this place is reserved for cognitive ability (e.g., Kuncel, Hezlett & Ones, 2001; Kuncel, Hezlett & Ones, 2004; e.g., Salgado, Anderson, Moscoso, Bertua & de Fruyt, 2003; Schmidt & Hunter, 1998). Nevertheless, there is empirical evidence for the incremental validity of personality in general and the BIG 5 specifically.

### *1.2.2. Use in Clinical Psychology*

Within the field of clinical psychology, the BIG 5 have been linked to many different psychological disorders ranging from alexithymia to schizophrenia. Some research results will be stated in the following paragraphs.

#### *1.2.2.1. Big 5 and Alexithymia*

The construct of alexithymia was introduced about three decades ago (Sifneos, 1973) and refers to a cognitive-affective disturbance manifesting itself with markedly reduced symbolic thinking, impoverished fantasy life, and limited ability to identify and verbally express emotions. Alexithymic individuals are characterized

by reduced ability to recognize and describe feelings, scarceness of fantasies, and a concrete cognitive style (Taylor & Bagby, 2004). Throughout the research history of the construct its stability and construct validity have been called into question. In a recent article Picardi, Toni, and Caroppo (2005) used a Big 5 measure to demonstrate stability and construct validity of alexithymia. The results confirmed stability. Moreover, and of more interest for the present project the correlations with neuroticism ( $r = -.42$ ) and openness ( $r = -.36$ ) were significant which indicates slight overlaps between alexithymia and personality as assessed by the Big 5.

#### 1.2.2.2. Big 5 and Depression

Depression is one of the most common diagnoses in contemporary psychotherapy. Some researchers have claimed that neuroticism as a trait is a good predictor for a major depression episode. However, results regarding this hypothesis were mixed. Schmitz, Kugler, and Rollnik (2003) used data from the National Comorbidity Study (NCS) to investigate this claim in a representative community sample. Results from a logistic regression revealed a strong interaction between neuroticism and self esteem in the prediction of a major depression episode. However, the strongest regression coefficient occurred for neuroticism. Thus, low emotional stability increases the odds of a major depression. Moreover, low self esteem in combination with low emotional stability also increases the odds of a major depression.

### 1.2.2.3. Big 5 and Eating Disorder

Anorexia nervosa can be separated into two subtypes, the restricting type (AN – RT) and the binge – eating/purging type (AN – BP) (Dacosta & Halmi, 1992). Bollen and Wojciechowski (2004) investigated personality differences between those two types. They used the NEO – FFI and found that participants belonging to the restrictive type scored significantly higher in agreeableness and conscientiousness compared with both other groups. The effect sizes were moderate. Taken together, both groups differed significantly from a control group in all personality factors except for agreeableness. The patient group achieved higher scores in neuroticism but lower scores in all other factors.

### 1.2.2.4. Big 5 and Schizophrenia

Within the field of research in severe psychosis and schizophrenia Big 5 measures have not been used as often as with other disorders. However, there are a few studies which applied Big 5 measures to schizophrenic patients.

Gurrera, Nestor, O'Donnell, Rosenberg, and McCarley (2005) investigated the relationship between cognitive and motor performance on the one side and the Big 5 on the other side within a patient and a control sample. A comparison of the two samples revealed significant and moderate to large differences in three of the five factors. Patients scored higher in neuroticism and lower in extraversion and agreeableness. However, when the performance in the cognitive and motor tasks and demographic variables were controlled the two groups no longer differed in any of the personality factor. Thus, the authors concluded that personality dysfunction

in schizophrenia may be mediated by disease-related changes in cognitive operations, or the neural processes underlying them.

#### 1.2.2.5. Big 5 and Personality Disorders

Obviously, the major field of clinical research using Big 5 measures is the field of personality disorders. Researchers usually try to connect personality factors to certain personality disorders (Blackburn, Logan, Renwick & Donnelly, 2005; Moxnes, 1996).

To give a short overview of some of the results a recent study will be reported (Furnham & Crump, 2005). Furnham and Crump used a Big 5 measure and a personality disorder measure in a sample of  $N = 431$  assessment center candidates. Even though the sample was not a clinical sample, the results confirmed prior outcomes of other studies. Summarizing, it can be said that neuroticism correlated with borderline and avoidant personality disorders, introversion correlated positively with avoidant and schizoid personality disorders, and negatively with histrionic personality disorder. Moreover, openness correlated with schizotypal and conscientiousness with obsessive – compulsive personality disorders. The correlations were small to moderate. All in all, the results were in line with meta – analytical findings (Saulsman & Page, 2004).

Saulsman and Page concluded in an earlier article that future research should “focus on the ability of neuroticism and agreeableness to screen for general personality disturbance and capture those qualities that should distinguish normal from disordered personality traits” (Saulsman & Page, 2003, p.83).

Summarizing the findings presented above, it is clear that the Big 5 also play an important role in contemporary clinical research. Especially neuroticism seems to be linked to a number of different disorders.

### *1.2.3. Use in Pedagogic Psychology*

One of the most interesting applications of the Big 5 within the field of pedagogic psychology is within the scope of academic emotions. Usually, research in this area was mainly concerned with test anxiety. However, a group around Pekrun broadened that perspective and also included other emotions (enjoyment, hope, pride, relief, anger, anxiety, shame, hopelessness, and boredom) into their research program (e.g., Pekrun, Goetz, Titz & Perry, 2002). Here personality traits, mainly trait anxiety, were used to demonstrate the discriminant validity of a specific test emotion questionnaire (Pekrun, Goetz, Perry, Kramer, Hochstadt & Molfenter, 2004).

The Big 5 were also used to investigate the implicit personality theories of teachers in judging their students (ter Laak, DeGoede & Brugman, 2001). Ter Laak, DeGoede, and Brugman analyzed 87 students ratings by four teachers. The resulting factor structure corresponded well with the Big 5. Thus, the authors concluded that teachers implicitly use a Big 5 – like personality theory to judge their students. Unfortunately, another result of that study was that the accuracy of these ratings is not very high.

Summarizing this brief excursion into the field of pedagogic psychology it is reasonable to conclude that the Big 5 have enriched research in this area as well.

#### 1.2.4. *Summary*

The cited studies from such various fields as clinical, organizational, and pedagogic psychology demonstrate that the Big 5 are an essential part of contemporary psychological research. Numerous studies have proven the value in predicting performance, psychological disorders, and theory building.

If a construct has impacted the scientific community as intensely as the Big 5 have done, it is essential that the means of measuring the construct have only minor flaws. However, a number of critics have not only doubted the construct itself but also the validity of questionnaires routinely applied to assess the Big 5. The most common criticisms refer to the number of factors, the methodology used to find the construct, the orthogonality of the five factors, the amount of trait and state variance, and finally the impact of social desirability. These typical criticisms are presented in the next paragraphs.

### 1.3. Concerns about the BIG 5

#### 1.3.1. *Number of Factors*

The number of factors has been debated for a long time already (Costa & McCrae, 1992a; Eysenck, 1992). As indicated in the beginning the number of factors deemed to be necessary to describe covariances in the self descriptions of people was not five when research started. Cattell for example proposed twelve to sixteen factors (Cattell, 1943a; Cattell, 1943b). This number was criticized as being too large from the day of the first publication. Later, Eysenck and Eysenck (1964) developed a questionnaire to assess what he called the Big 3: Psychoticism,

Extraversion, and Neuroticism. Both latter traits are still part of the Big 5 model known today.

However, the character of psychoticism raised many discussions. The last publicized controversy between Eysenck on the one hand and Costa and McCrae on the other hand was published in 1992 (Costa & McCrae, 1992a; Eysenck, 1992). The conclusion from that dispute was that Costa and McCrae succeeded in convincing the majority of the scientific community that psychoticism is nothing but "... a combination of low Agreeableness (A) and low C ..." (Costa & McCrae, 1992a, p. 861).

Costa and McCrae originally started their research endeavor with the Big 2, namely neuroticism and extraversion (John, et al., 1988). Later they incorporated openness into their concept and decided to research the Big 3 (Costa & McCrae, 1980). However, due to the evidence for two further factors, Costa and McCrae again expanded their model and included agreeableness and conscientiousness (Costa & McCrae, 1992b). This model was welcomed warmly by the scientific community (see above) and has proven its value.

However, in recent years the debate concerning the number of factors has risen again (Ashton, Lee, Perugini, Szarota, de-Vries, Di-Blas, Boies & De-Raad, 2004; Saucier, 2003; Saucier, Georgiades, Tsaousis & Goldberg, 2005). Saucier et al. (2005) developed a personality questionnaire for Greeks. Results showed that a six factor solution might be sensible. Yet, in a different paper Saucier (2003) found that one and two factor solutions were the most replicable solutions.

The different number of factors reported for different samples and with different adjective lists might be prove for the repeatedly expressed criticism concerning exploratory factor analysis as the method of choice in research engaged with the question of the number of factors. This issue will be described next.

### 1.3.2. *Methodological Issues*

The use of exploratory factor analyses in psychological research is wide spread. Exploratory factor analyses search for latent factors which are responsible for the covariations between variables. The method was successfully used in intelligence research (Buehner, Krumm & Pick, 2005; Buehner, Ziegler, Krumm & Schmidt Atzert, 2006; Carroll, 1993).

With Cattell exploratory factor analysis found its way into personality research. Due to the higher efficiency of modern statistical software packages factor analyses today can easily be conducted with large numbers of items and participants. Costa and McCrae also use exploratory factor analyses to conduct much of their research. This approach, however, has been criticized because of the many subjective decisions made in the course of exploratory factor analyses.

Block (1995) is one of the most active antagonists of the Big 5. Within his paper from 1995 he also criticized that factor analysis is always bound to rely on a number of subjective decisions such as extraction criteria, rotation, and extraction method. Moreover, he stated that the homogeneity of the sample affects the results derived in a factor analyses. Besides this, Block also argued that exploratory factor analysis only uses the common variance of the items. He reasoned that items which correlate only lowly with the other items might still be reliable and therefore indices of personality aspects. However, these indices would not be considered in a factor analytical approach since only items with high loadings on the factors were kept. Thus, according to Block, the mere application of factor analyses might lead to a loss of important information regarding the personality of humans.

Costa and McCrae (1995) replied to the critique by Block. Their, in parts rather lofty, reply was mostly concerned with listing empirical evidence for the merits of

the Big 5. Concerning the use of exploratory factor analyses their reply was short. They simply stated that they agree with Block and also wish that such techniques were not applied as senseless as in some cases. Moreover, they also called for more sophisticated methods to analyze self report data.

### *1.3.3. Orthogonality*

Throughout the history of factor analytical research in personality researchers have always preferred a Varimax rotation in their proceeding. The main reason for that was that orthogonal solutions are easier to interpret. The goal of a rotation, may it be orthogonal or oblique, is to achieve simple structure. That means, items should have high loadings on one factor and low loadings on the other factors. However, the choice of the rotation method should not be based on convenience. Costa and McCrae preferred orthogonal rotation because it provides more parsimonious solutions (1995, p. 218).

Mounting empirical evidence, however, has accumulated demonstrating that there are non trivial correlations between some of the personality factors. Mount, Barrick, Scullen, and Rounds (2005) collectively analyzed these data within a meta – analysis. The results showed that there are substantial correlations between personality factors. As is the state of the art in meta – analysis these correlations were corrected for restrictions in range and reliability. Thus, Mount et al. always refer to true correlations. The largest true correlation occurred between neuroticism and conscientiousness and amounted to .50.

Costa and McCrae acknowledge these correlations and attribute them to a lack in discriminant validity for some of the facets. Furthermore, they advice users of their questionnaire to compute factor scores based on the results from the American

normative sample. However, within this sample the uncorrected correlation between neuroticism and conscientiousness was .53. This is even larger than the true correlation reported by Mount and colleagues. All this would not be a problem if the correlations among the personality factors did not raise doubts concerning the construct validity of the Big 5. Maybe personality factors are not uncorrelated after all?

Biesanz and West (2004) offered a different explanation for the correlations. They argued that some of the items have a positive connotation and some have a more negative connotation. Furthermore, they said that the intercorrelations between personality factors result from the use of only one information source. They conducted a Multi – Trait – Multi – Method analysis and found that the intercorrelations almost disappeared when several information sources were combined. This phenomenon was called evaluative bias.

While this result offers an explanation for the intercorrelations it is not satisfying. Usually, only information from one source is available. Moreover, the explanation does not offer any help to the user.

Within the present project a different approach to the problem was chosen. The main idea behind this approach were the state and social desirability impact on personality questionnaires. The state impact will be discussed in the next section before the impact of social desirability is introduced.

#### *1.3.4. Trait and State Variance*

The goal of most questionnaires and especially personality questionnaires is the assessment of traits. A trait can be defined as enduring characteristic or disposition (Liebert & Liebert, 1998, p. 184). The reason why scientists and practitioners alike

want to assess traits is that the prediction of behavior is more reliable and promising if the measured aspect is stable over time and situations. Thus, one of the goals in questionnaire construction is that the items tap aspects of traits.

However, there is a dilemma. People do not behave stable over different situations. This dilemma has become known as the consistency paradox. Obviously the situation also plays an important role in determining people's behavior. One of the first to point out this fact was Mischel (Mischel, 1968; Mischel & Peake, 1982). He observed that people see themselves as relatively stable personalities. In reality, however, the behavior of a person is not cross – situational consistent. Nevertheless, there is evidence for temporal consistency. This term refers to the fact that a person will behave relatively equal in a specific situation if the situation reoccurs. Mischel strongly believed that the situational influence cannot be neglected and interacts with personality traits to determine behavior. This interactionist view is accepted by most scientists today. Yet, when it comes to the prediction of behavior mostly traits are used since it is assumed that personality questionnaires assess only traits. This assumption, however, is not correct.

Besides traits, psychologists often refer to states when characterizing people. States can be regarded as temporary conditions (Liebert & Liebert, 1998, p. 185). A state that has attracted a lot of attention in research is state anxiety. Many studies were conducted in order to distinguish between trait and state anxiety and their impacts on different aspects of life. In 1992 Steyer, Ferring, and Schmitt (1992) proposed a model which acknowledges the simultaneous presence of state and trait in personality measures. Using structural equation modeling and at least two measurement points the variance of a personality questionnaire can be split into trait variance, state variance, interaction variance, and error variance. With the help of this study Deinzer and colleagues (Deinzer, Steyer, Eid, Notz, Schwenkmezger,

Ostendorf & Neubauer, 1995) demonstrated that up to twenty percent of the variance in a questionnaire filled out in ordinary situations can be attributed to state. In other words, the situational influences a person is exposed to impact the answer to a personality questionnaire.

Even though this result has been known in the scientific community for more than ten years, the consequences of this situational influence have not been investigated in depth. Consequently, the present study will explore situational influence on the construct and criterion validity of a personality questionnaire.

#### *1.3.5. Social Desirability and Its Impact on Construct and Criterion Validity*

One of the main concerns regarding the use of personality questionnaires is the impact of social desirability. Social desirability is viewed as a response set by most researchers (Murphy & Davidshofer, 2001). The problem regarding social desirability consists of the assumption that it increases or decreases scores in personality questionnaires according to the situation.

A prominent contributor to the body of research concerned with social desirability is Paulhus (e.g., Paulhus, 2002; Paulhus, Harms, Bruce & Lysy, 2003). Paulhus has dedicated most of his life as a scientist to the research of the construct of social desirability. As indicated above, social desirability is a form of deception since people do not portray themselves honestly. Paulhus distinguishes two types of deception. On the one hand there is self deception. This means that people distort their answers to a questionnaire in a direction which is close to the personal ideal of that person. This process is supposed to happen unconsciously. Therefore, some scientists have strictly tried to avoid the use of questionnaires. McClelland (1987, p. 187) for example once wrote that a scientist cannot believe what the people say

about themselves. However, of more interest is the second type of deception, the other deception.

Other deception, also called impression management, is meant when people intentionally endorse other rating categories in a questionnaire in order to impress others. In an application setting for example, applicants supposedly portray themselves as more conscientious and less neurotic than they actually are. This has also been called faking.

The role of faking or social desirability on the construct and the criterion validity of personality questionnaires has been researched broadly (e.g., Collins & Gleaves, 1998; Ellingson, Smith & Sackett, 2001; Pauls & Crost, 2004; Pauls & Crost, 2005b; e.g., Stark, Chernyshenko, Chan, Lee & Drasgow, 2001; Zickar & Robie, 1999a). One of the findings of these studies is that faking causes correlations between personality factors to increase. A result of these higher correlations was the extraction of a sixth factor. Schmit and Ryan (1993) extracted this factor with loadings from almost all personality facets and called it the ideal employee factor. Despite this evidence for the impact of social desirability on the construct validity there are some scientists who claim that no serious impact can be found. Ellingson, Smith, and Sackett (2001) reported that higher order factor structures for people high and low in social desirability do not differ. However, their study used social desirability questionnaires to select the samples. Meanwhile, there is evidence that such measures can be faked as well and as easily as other questionnaires (Pauls & Crost, 2004). Thus, the sample identification might have been flawed. In a different study Smith, Hanges, and Dickson (2001) compared the structure of the Big 5 in an applicant and a student sample and reported no differences in loading patterns. A closer look at their models shows however, that there are differences in the size of the intercorrelations between personality factors. Within the applicant sample these

intercorrelations were higher. Thus, the conclusion that construct validity is not affected by social desirability or faking might have been too early.

A different but related branch of research investigated the impact of social desirability or faking on the criterion validity of personality questionnaires. While there is a large amount of evidence which shows that mean scores of such measures increase dramatically (e.g., Rosse, Stecher, Miller & Levin, 1998; Ziegler, Bühner, Krumm & Schmidt Atzert, submitted-a) due to faking there is also evidence that the criterion validity, at least as far as the overall effect size is concerned, is not affected by faking. Ones and Viswesvaran in collaboration with other researchers (Ones, Viswesvaran & Reiss, 1996; Viswesvaran & Ones, 1999; Viswesvaran, Ones & Hough, 2001) used meta – analytical techniques to show that even though mean scores increase criterion validity as effect size is not impacted. This is an encouraging finding for all people interested in using personality questionnaires. However, there is also evidence for an indirect impact of faking on such questionnaires. Hough (1998) reported that correcting for faking strongly changes the ranking of people. While this again had no influence on the criterion validity it underlines that personality questionnaires no longer only assess personality when they are faked. This aspect will also be investigated in the present project.

#### *1.3.6. Summary*

The description of the different critical aspects of the Big 5 shows that especially the construct validity and the criterion validity of questionnaires measuring the construct have been questioned. Within the present project three studies were conducted to find answers to these questions. However, there is one aspect which governs all research attempts undertaken in these three studies. From the studies

by Deiner et al. and Steyer et al. it is clear that every personality questionnaire not only assesses trait but also state variance. This state variance, or as it will also be called from here on situational demand, might be responsible for some of the critical aspects of the Big 5. For example, it is reasonable to assume that situational demand is part of the orthogonality problem and therefore impacting construct validity. Moreover, the research cited above also shows that faking, at least indirectly, also impacts criterion validity. Faking can be understood as situational demand. If one wants to distort a questionnaire, the most likely reason is the situational demand the person feels. Consequently, the present project also investigated the impact of situational demand on criterion validity.

Most of the research concerned with the Big 5 or faking uses complex methodological approaches to tackle the research questions. However, so far little is known about the actual psychological process occurring when people fake. In other words, it is unclear which strategies people might use and why they do so. These questions will also be part of this research project.

Finally, many efforts have been conducted to guard against faking. The present project will also test a new method, namely Mixed Rasch models, to detect faking and eliminate its impact on scores.

Within the following paragraphs these research goals and how they are implemented in the studies will be explained in more detail.

## 1.4. Goals of the present project

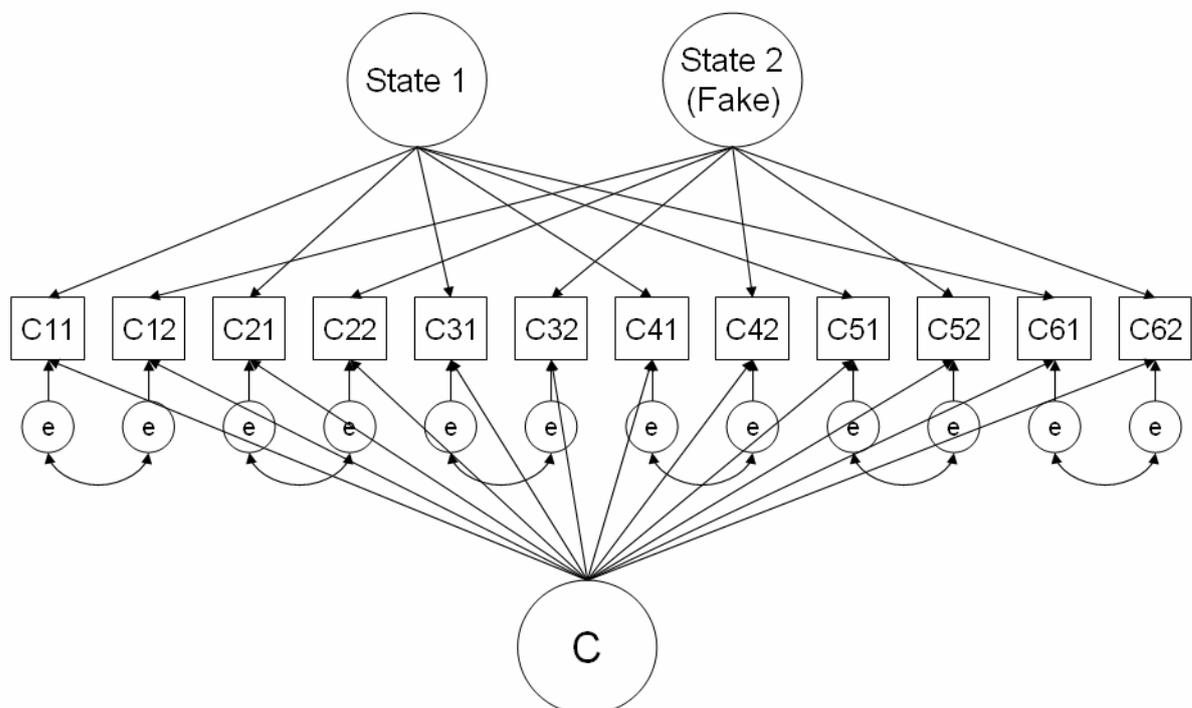
### 1.4.1. *Impact of Situational Demand on the Construct Validity of the BIG 5*

As has been described above the construct validity of the Big 5 has been questioned by some researchers. From all the critical points mentioned above the reported correlations between theoretically uncorrelated personality factors will be an issue in the following studies. The results by Deiner et al. (1995) demonstrated that situational influence can account for up to twenty percent of the systematic variance in a personality questionnaire. Moreover, other results (e.g., Pauls & Crost, 2005b) showed that the correlations between personality factors increase dramatically when situational demand and thus, situational influence increases. Drawing on these results the following hypothesis was tested in this project: Situational demand can be regarded as a systematic measurement error which adds on to the systematic variance. Since this error variance affects all faked variables, correlations between the variables increase. This means that situational demand causes the intercorrelations between personality factors and controlling for situational demand will yield uncorrelated personality factors. This hypothesis was investigated in study 1 and study 2 (see chapter 2 and 3).

In order to test this hypothesis, it is necessary to separate trait and state variance. The proposed model by Steyer et al. (1992) was used to do that. This model uses at least two measurement points to separate the different variance shares. The idea is that the trait variance should be part of every measurement point. In other words, whenever I ask a person about her or his openness to experience the answer should always contain trait variance. However, each time a person is questioned certain situational circumstances also influence the answer.

Yet, in comparison with trait variance this influence is bound to the specific situation. This means that the state variance might impact different variables at different times. For example, when a person has been studying for an important exam for weeks and therefore has not been to parties at all, her or his extraversion scores will be different compared with another time when she or he has just returned from a two week vacation in Ibiza. Consequently, state variance is only drawn from the variables measured at the same time. Since the items used to question a person are the same at each measurement point, the error variances of identical facets are allowed to correlate. The resulting structural equation model, also called Latent State Trait Model (LST), is exemplified for the conscientiousness factor in figure 1.1.

Figure 1.1 Example of a Latent State Trait Model



It can be seen that the facets C1 (competence) to C6 (liberation) have been measured twice as implied by the second number. The trait variance (C) is then

extracted at one side and draws variance from all facets. State variance exists for both measurement points, therefore, two latent state variables exist. These variables only draw variance from facets which have been assessed at the same time. For example, state 1 only has loadings from facets with a 1 as second digit. The correlated error variances constitute the fact that the same facet has always been measured with the same items. The figure also demonstrates one of the main ideas of the present project. If the situational demand at one measurement point is simulated to represent an applicant setting, the state variance can be regarded as faking variance. In other words, if participants are given a specific instruction to fake, then the state variance represents faking variance. Since the trait variance can still be drawn because of the first measurement point this faking variance no longer contains trait variance. Thus, a separation of trait and faking variance can be achieved with this design. It was tested whether controlling for situational demand would yield uncorrelated personality factors as was hypothesized.

The same question was also looked at in study 1 (chapter 2). However, in this study no second measurement time existed. Thus, only aspects of LST models were realized. Strictly speaking, a bimodal model with a latent trait and a latent state variable simultaneously drawing variance from the variables was constructed. However, the state variance has loadings from all used variables since it was assumed that the situational influence impacts all variables simultaneously. The trait variances, however, only have loadings from their respective facet variables. Thus, state and trait variances were also separated in study 1. The aim of study 1 was to find out whether situational demand was responsible for differing correlation patterns between the personality factors in different cognitive ability groups.

Summing up, study 1 and study 2 used aspects of or a complete LST design to separate state and trait variance and to investigate how controlling for situational demand influences intercorrelations between traits.

#### *1.4.2. Impact of Situational Demand on the Criterion Validity of the BIG 5*

In the first part of the introduction research results regarding social desirability and its impact on the criterion validity of personality questionnaires have been illustrated. The prevailing conclusion of these studies is that social desirability does not influence criterion validity.

In studies 2 (chapter 3) and 3 (chapter 4) this conclusion was scrutinized. The model described in the previous section allowed the separation of state and trait variance. In order to study the effects of situational demand on the criterion validity of the Big 5, the LST model was combined with an Individual Causal Effect Model (Steyer, 2005). Individual Causal Effect Models (ICE) use a control and an experimental group. In study 2 the control group had to fill out the personality questionnaire twice, each time as honestly as possible. The experimental group, however, only answered honestly the first time. The second time this group was given a specific faking instruction. With the ICE design any changes occurring in the experimental group at time 2 which do not occur in the control group can be attributed causally to the faking instruction. Thus, it is possible to find out whether personality factors still differ when faking is controlled. Moreover, and more interesting, it is possible to regress a criterion on the personality factors and the faking factor. Thus, the question where the predictive power of faked personality questionnaires stems from can be answered.

In study 3 (chapter 4) a different methodological approach was chosen. Applying Mixed Rasch Models (MRM, Rost, 1991) different faking styles were identified. Mixed Rasch Models are a combination of Latent Class Models and Rasch Models. Such models can be applied to questionnaire or test data. Instead of analyzing variance these models analyze answer patterns, though. Within Latent Class models the most likely number of different classes necessary to describe the answer patterns is computed. People are assigned to one of these qualitatively different classes. However, within each class every person is assumed to have the same probability of solving an item in an intelligence test or endorsing a category in a questionnaire. Rasch models on the other hand assume that the answer patterns observed can be attributed to individual differences in one and the same latent variable. In other words, a Rasch model tests whether it is likely that one dimension is responsible for the observed differences in answer patterns. In contrast to Latent Class models people can differ quantitatively in this latent variable. MRMs combine both theories and look for the number of qualitatively different classes in which the Rasch model holds true and which are most likely given the observed answer patterns. The classes still differ qualitatively. Members of one class, however, also differ in their probability of answering a question. Thus, MRMs are comparable with cluster analyses. However, cluster analyses use distances to group people. Therefore, people with the same total score (number of solved items or sum of endorsed categories) will most likely end up in the same category. However, they might have totally different answer patterns. For example, two people might have solved ten items in an intelligence test. The first person solved the first ten and the second person the last ten items. Is it really likely that both people belong to the same cluster? Since MRMs analyze answer patterns it is possible that people with

the same number of solved items will be assigned to different classes according to their answer patterns. This is especially interesting in the research of faking.

Summing up, study 3 used MRMs to find different faking classes and then computed the criterion validity of honest, faked, and MRM corrected (see 1.4.4.) scores. Applying this technique, a more differentiated look at the impact of situational demand on criterion validity is possible.

#### *1.4.3. The Psychological Process behind Situational Demand*

Part of study 3 was also a qualitative study. Such studies are relatively rare in contemporary psychology since they do not offer sophisticated methodological tests to guard against chance occurrences. In early psychological research studies, for example Freud (e.g., Freud, 1910), case studies were used more frequently and in some instances exclusively. Nowadays, qualitative studies are used to improve the legibility of questionnaires for example (e.g., Dillman & Redline, 2004; e.g., Willis, 2004). Usually, cognitive interviews are applied here. In these cognitive interviews participants either express any thoughts they have during the work on the questionnaire (concurrent cognitive interview) or they are interviewed afterwards (retrospective cognitive interview). In both cases the aim is to find difficulties in item wording or categories which impede answering the questionnaire. In psychotherapy such think aloud techniques are used to explore semantic networks of fear. However, when it comes to faking and the specific psychological processes happening when people intentionally distort their questionnaire answers, qualitative analyses have not been used before.

Therefore, study 3 used both cognitive interview techniques to explore this psychological process. Two experts analyzed the thought protocols and interview

results in order to identify different faking strategies. Moreover, the exact thought process was analyzed for systematic patterns.

#### *1.4.4. Dealing with Situational Demand*

Researchers and practitioners alike have tried for a long time now to find a way to guard against faking or to eliminate its effects from questionnaire answers. In the beginning mostly so called lie or faking scales were developed (e.g., Eysenck & Eysenck, 1964). Especially the Marlowe – Crowne scale has been used by many researchers to spot faking (Crowne & Marlowe, 1960). Usually, a questionnaire is developed and its correlation with the Marlowe – Crowne scale or another social desirability scale is determined. If the correlation turns out to be low, it is concluded that the questionnaire is not susceptible to faking (e.g., Sokolowski, Schmalt, Langens & Puca, 2000). Results from faking studies indicate that this conclusion might be wrong, though. Ziegler, Bühner, Krumm, and Schmidt – Atzert (submitted a) reported that such a measure is also susceptible to faking despite the low correlation to the Marlowe – Crown scale reported by the test author. Moreover, Pauls and Crost (2005) demonstrated that impression management and deception scales can also be faked in an applicant setting. Nevertheless, such scales have proven their value in identifying fakers (e.g., Baer, Wetter, Nichols, Greene & et al., 1995). This task also seems to be of more interest to most practitioners. Be that as it may, identifying a faker raises two questions. What do I do with a faker? Is it really a faker or does the person really have such a highly developed trait? Other tactics have been developed to guard against faking or to detect faking. Ipsative measures were long believed to be the best way to help against faking (e.g., Baron, 1996; Bowen, Martin & Hunt, 2002; Martin, Bowen & Hunt, 2002). However, they have

two disadvantages. First of all, they only allow finding out personal preferences but do not allow interindividual comparisons regarding the specific amount of a trait. In other words they help to find out whether a person is more extraverted than conscientious but not whether the person is more extraverted than another person. Secondly, these measures can be faked as well. Ziegler et al. discussed all these methods and added cut off scores to the list of possible aids helping to guard against faking. However, they concluded that objective measures (McClelland, Koestner & Weinberger, 1989) might be the best way to guard against faking.

Within study 3 a new approach to the problem was chosen. As mentioned above, study 3 used Mixed Rasch Models (MRM) to find different faking classes. MRMs do not only allow finding different classes, they also estimate person parameters for each participant according to the assigned class. In other words, the estimated trait score is corrected for the response set chosen. Rost, Carstensen, and von Davier (1997) could show that under ordinary circumstances two response styles for answering a Big 5 questionnaire can be identified. There are people who prefer extreme categories and people who prefer middle categories. When the person parameter is estimated in a MRM, these preferences, which influence the score but are not due to trait differences, are corrected. Thus, middle crosser get slightly larger and extreme crosser slightly smaller person parameters. Thus, with the help of MRMs it might be possible to correct for faking and reconstruct parameters close to the honest scores. This was tested in study 3.

#### *1.4.5. Summary and outlook*

In the following chapters the three different studies undertaken to pursue the four goals just stated will be described. The first study (chapter 2) analyzed data collected by Toomela (2003) to explore the impact of situational demand on the construct validity of a Big 5 questionnaire in different ability groups. In study 2, an experimental design with two groups and two measurement points was realized to separate trait and faking variance. Thus, the impact of situational demand on construct as well as criterion validity could be assessed. Finally, within the third study results from a qualitative analysis regarding different faking styles are reported. Moreover, MRMs were used to explore differences between different faking classes.

After the three studies have been described and discussed, chapter 5 will shortly summarize the results and then provide a discussion of the results from all three studies regarding the four research goals.

**Chapter 2: A Reanalysis of Toomela (2003): Personality Structure - A  
Question of IQ or Situational Demand?**

## **Chapter 2. A Reanalysis of Toomela (2003): Personality Structure - A Question of IQ or Situational Demand?**

The BIG 5 personality model is probably the most sophisticated personality model in psychology. The first to use the name BIG 5 for the personality dimensions which consistently occurred in different samples was Goldberg (see Goldberg, 1981). The five factor model (FFM) as we know it today, consisting of five higher order personality traits (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), was proposed by Costa and McCrae (1992b). Meanwhile extensive research has been conducted and the Big 5 have been replicated in different cultures and with different instruments. The BIG 5 can be understood as five independent dimensions which describe the pattern of covariations between individual personality trait descriptions with a sufficient accuracy. Therefore, the most common method to prove the structure of BIG 5 questionnaires is a Varimax rotated factor analysis (e.g. Toomela, 2003; Costa & McCrae, 1992b).

However, there also seems to be a lot of empirical evidence that the BIG 5 are not as uncorrelated as proposed. In a recent meta – analysis by Mount, Barrick, Scullen, and Rounds (2005) substantial true correlations between the BIG 5 were reported. Most notably for this paper is the reported true correlation of .50 between neuroticism and conscientiousness. This is especially interesting since Toomela (2003) could not clearly distinguish these two dimensions in a group of highly intelligent men using a Varimax rotated factor analysis. In his paper he reported difficulties in fitting the FFM in a group of males with above average intelligence. In this particular group, a four factor model seemed to be the better solution.

Meanwhile, the data have been reanalyzed by Allik and McCrae (2004). Those authors followed recommendations by McCrae, Zonderman, Costa, Bond, and Paunonen (1996) to use a confirmatory approach to test factor structure after an exploratory approach. Using an orthogonal Procrustes rotation, Allik and McCrae found that Toomela's data could be fitted to the American normative sample quite well. Therefore, one could argue that the problem was solved. However, we want to approach the topic from a different angle. Costa and McCrae (1992b) reported a correlation of  $r = -.53$  between neuroticism and conscientiousness in the normative sample which was used as target in the Procrustes rotation. Thus, even though the factor solution shows convergence with another factor solution, the question of where the overlap between the factors stems from remains. This question has been discussed elsewhere already. Costa and McCrae (1992) commented on this and recommended the use of factor scores to circumvent the problem. Biesanz and West (2004) claimed that evaluative bias was responsible for the overlap. They used a multi – trait – multi – method approach and found that the intercorrelations between personality factors decrease when different information sources are used.

Within this reanalysis we want to use structural equation modelling to show that situational demand increased the state variance in specific scores and caused the inflated correlation between neuroticism and conscientiousness. The starting point for this claim was the increased correlation between neuroticism and conscientiousness reported by Toomela. Moreover, from the study by Pauls and Crost (2005b) it is known that correlations between personality factors increase due to situational demand (faking). Pauls and Crost (2005b) could show that the correlations between personality traits increase with situational demand (selection scenario). This effect only occurred for the traits the subjects believed to be job-relevant. The authors reported a correlation of  $r = -.76$  between conscientiousness

and neuroticism, exactly the two personality dimensions Toomela could not distinguish clearly. The result that the structure of the BIG 5 might change in applicant settings has already been reported by Schmit & Ryan who found a sixth factor which they called ideal employee factor (Schmit & Ryan, 1993). However, this factor has not always been replicated (see also Cellar, Miller, Doverspike & Klawnsky, 1996; De Fruyt, Aluja, Garcia, Rolland & Jung, 2006).

This alone would not suffice to start a reanalysis of the Toomela data since participants here did not have specific faking instructions. However, Zickar, Gibby, and Robie (2004) used Mixed Rasch Models and found that even in incumbent samples where faking was always believed not to happen, a sizeable amount of people intentionally distorted their answers. The authors concluded that “the assumption that applicants fake and incumbents do not appears to be untenable ...” (p.186). Thus, it is reasonable to assume that the same thing happened in the Toomela data and caused the inflated correlation. This was tested in the course of this reanalysis.

The first step in the reanalysis was to investigate the question why situational demand was not felt equally by all participants. Since the differences in personality structure were found in groups of different intelligence, cognitive ability is a likely candidate for this phenomenon. The relationship between personality and intelligence has received a lot of attention in the last decade (e.g., Ackerman, 1996; Chamorro Premuzic & Furnham, 2003; Chamorro Premuzic & Furnham, 2004; Chamorro Premuzic, Furnham & Moutafi, 2004; Furnham, 2002; Furnham & Chamorro-Premuzic, 2004; Furnham, et al., 2002; Furnham, Moutafi & Chamorro Premuzic, 2005; Toomela, 2003). Even though the results repeatedly show significant correlations between different personality traits and intelligence (Harris, Vernon & Jang, 2005; Moutafi, Furnham & Crump, 2003; Moutafi, Furnham &

Paltiel, 2004; Zeidner & Matthews, 2000), these correlations are rather low. Moreover, drawing from the differentiation hypothesis from intelligence research (Deary & Pagliari, 1991; Spearman, 1927), some researchers have assumed that the variability in personality traits also increases with the level of intelligence (Brand, Egan & Deary, 1993; Harris, et al., 2005). This would mean that the BIG 5 structure should actually emerge easier in groups of highly able people. This is also in line with recent findings by Möttus, Allik, and Pullmann (2006) who also used an Estonian sample. Those authors reported only small differences in personality structure between a group of cognitive able and a group of cognitive less able participants. They also reported that intercorrelations between personality dimensions were somewhat smaller within the more able group. However, in the study by Toomela these results could not be observed.

To sum up this first part, there are two differing points of view: (1) The BIG 5 are a structure of uncorrelated personality dimensions, and (2) the BIG 5 are a structure of correlated personality dimensions. The latter allows phenomena like the increased correlation reported by Toomela. The first view does not allow such correlations.

Using a different statistical method, we want to bring a new perspective into this discussion. At least since Mischel (Mischel, 1968; Mischel & Peake, 1982), there has been debate about the situational influence on behavior and its interaction with personality. Today, nobody would disagree that actual behavior is determined by the situation and personality traits. The controversy remains on the amount of the situational influence. In the 1990's the debate was refueled with a new methodological approach called Latent-State-Trait Theory (LST) (see Steyer, et al., 1992; Steyer, Schmitt & Eid, 1999). This theory pays attention to the fact that states might also influence the measurement of a trait. The idea is that each measure of

personality contains trait variance, state variance, error variance and variance due to the interaction of trait and state. In a large scale research project conducted by Deinzer, Steyer, Eid, Notz, Schwenkmezger, Ostendorf, and Neubauer (1995), three commonly used personality questionnaires were analyzed to see how much trait variance they actually contain. The three questionnaires applied were the NEO Five-Factor Inventory (Costa & McCrae, 1992b), the Eysenck Personality Inventory (Eysenck & Eysenck, 1964) and a well known German personality questionnaire, the Freiburger Personality Inventory (Fahrenberg, Hampel & Selg, 1984). The results of the analyses conducted by the authors show that there is a substantial amount of trait variance. However, they could also show that the amount of variance due to state or interaction can amount to as much as 20% of the total variance. Unfortunately, it is rather difficult to compute the different variance amounts. First of all, the questionnaire has to be administered at least twice. Secondly, the computation is done by structural equation modeling which requires large samples for the kind of model applied in this case. These may be reasons why the LST approach has not been followed up to a greater extent. However, the results reported here clearly indicate that the situational influence on the measurement of personality should not be neglected. As Deinzer et al. (1995, p.7) formulated pointedly: "... we always measure persons in situations, not persons; there is no psychological measurement in the situational vacuum".

All in all, there are differing views on the interconnectedness of the BIG 5 and the situational influence on the general structure. Our starting point in trying to further explore these points of views is the reanalysis of data reported by Toomela. The goal of the present study is to propose an extension to the traditional FFM which pays attention to situational demands. By doing this, we also hope to shed some light into the structure of the BIG 5: correlated or not? Before the current

reanalysis will be explained further, a short overview over the study by Toomela is given.

### 2.1. The study by Toomela

In the study conducted by Toomela  $N = 912$  native-born Estonian men aged 17 to 68 years participated and filled out the Estonian NEO-PI-R (Kallasmaa, Allik, Realo & McCrae, 2000), a structure of word meaning test, and the Estonian Cognitive Ability Scale (Pulver, 1999). The analyzed sample, which is also the basis for this paper, contains  $N = 870$  people. Five cognitive ability as well as five structure of word meaning groups were built. The cognitive ability groups did not only differ in their mean level of personality traits but also in the number of personality factors. The contra intuitive result was that within the group with the highest cognitive ability only 3 or 4 factors emerged using exploratory factor analysis. This would be contradictory to the differentiation hypothesis. Moreover, the amount of variance explained by a FFM decreased in this group after steadily increasing from group to group before. For the word meaning structure groups the results were clearer. The FFM was more coherent and explained more variance with increasing word meaning structure. Toomela draws on these results to conclude that a cultural factor (word meaning structure) offers important tools (words) for the development of personality, and that intelligence can be understood as the mastery of these tools.

An alternative explanation for the smaller number of factors in the most able group might be the situational demand. This assumption will be supported in the following section. All participants (except for  $N = 21$  students) were part of the Estonian military or somehow affiliated with it. The sample contains recruits ( $N =$

400), non-commissioned officers ( $N = 201$ ), commissioned officers ( $N = 186$ ), and members of the Estonian Defense League ( $N = 62$ ). It seems reasonable that this fact might have influenced the data.

## 2.2. Reanalysis

Drawing on the results by Zickar, Gibby, and Robie (2004), Schmit and Ryan (1993) as well as Pauls and Crost (2005b), we assume that the critical results found by Toomela might be due to situational demand. As mentioned above, Zickar, Gibby, and Robie demonstrated that faking even occurs in incumbent samples and Schmit and Ryan could show that the BIG 5 model does not fit in an applied setting while a six factor model does. Pauls and Crost's results were derived in a faking experiment and show that participants do not fake all personality dimensions but only those they believed to be job-relevant. The correlations between those factors increased dramatically. We propose that the participants with the highest cognitive ability felt a situational demand and "faked" their answers. A recent paper by Konstabel, Aavik, and Allik (in press) lends further support to this hypothesis. Those authors report that the highest correlations between social desirability measures and BIG 5 dimensions in an Estonian sample in an applicant setting occurred for conscientiousness and neuroticism.

Thus, the hypothesis raised here that not intelligence but situational demand led to an increased correlation between conscientiousness and neuroticism finds additional support. However, this assumption raises the question, why would cognitive able people fake? To answer this question, we took a closer look at the original data and found that the higher the cognitive ability, the higher the military rank. Table 2.1 illustrates this fact.

Table 2.1 Percentage of military rank in each cognitive ability group

	Total N	CA1	CA2	CA3	CA4	CA5	<i>M</i>	<i>SD</i>
Recruits	400	28.5	25.3	24.0	12.0	10.3	16.8	7.3
Students	21	14.3	28.6	23.8	19.0	14.3	20.1	7.7
noncommissioned officers	201	14.4	17.4	30.8	21.4	15.9	21.7	6.5
Officers	186	5.4	12.4	30.1	29.0	23.1	24.6	5.8
EDL	62	6.5	12.9	21.0	17.7	41.9	27.6	5.6

Annotations. EDL stands for Estonian Defense League. CA1 to CA5 represent the 5 ability groups: CA1 (1–12 correct answers), CA2 (13–18 correct answers), CA3 (19–24 correct answers), CA4 (25–28 correct answers), and CA5 (29–35 correct answers) according to Toomela (2003). The maximum number of points was 36.

Therefore, we conclude that a situational demand was not felt because of a higher cognitive ability, but rather because of a higher military rank. This means that people with a higher military rank would have answered the questionnaire in a way they believed to be typical or representative for a military person. To put it short, people with a higher military rank felt a stronger situational demand. We assume that this demand influenced their answer style. Results of this are answers believed to be prototypical or representative for military persons. This would lead to increased correlations between the factors assumed to be prototypical since they would also contain additional variance due to the situation. Thus, it would be hard to separate these factors, which in turn would explain the smaller number of factors found by Toomela. Of course this assumption is speculative. Participants were informed that the data would be handled anonymously and only be used for research. Maybe the latter aspect motivated high ranking men to portray a more prototypical military personality in order to make the military as a whole look better. For all the following analyses we compare three military groups: recruits, non-commissioned officers, and officers. We collapsed students and recruits into one

group. Both have the smallest scores in cognitive ability and probably the least motivation to appear as a “prototypical” military person. We assume this because the former were not actually affiliated with the military and the latter were recruits doing their obligatory military time in the Estonian army. We also collapsed the officers and the members of the Estonian Defense League. Both groups have the highest results in the cognitive ability test and probably the highest motivation to be a representative military person because the military is a big part of their daily life.

The fact that Toomela found high cross loadings for conscientiousness facets on neuroticism and vice versa is in line with findings by Pauls & Crost (2005b) who found that people only fake personality dimensions believed to be job relevant, leading to increased correlations between those dimensions (see also Konstabel, Aavik & Allik, in press). Looking at the results found by Toomela reveals that conscientiousness and neuroticism have a bivariate correlation of  $r = -.77$  ( $p < .01$ ). This clearly exceeds the true correlation reported by Mount et al. (2005) and also the correlation of  $-.53$  reported by Costa and McCrae (1992). Moreover, the correlation reported by Mount et al. was corrected for attenuation as well as range restriction which the bivariate correlation is not. To provide evidence for the hypothesis that not intelligence, but situational demand caused by different military status caused this correlation, the size of it should be equally high in the group containing officers and members of the Estonian Defense League. Correlation analysis yields a result of  $r = -.67$  ( $p < .01$ ). This correlation is only slightly smaller than in the highest cognitive ability group. Thus, a first shred of evidence for the hypothesis is provided.

Another prerequisite for further analysis is that the factor structure in the three military groups reveals the same peculiarities as in the cognitive ability groups. Results from Varimax rotated Principal Component analyses are in table 2.2.

It can be seen that the number of substantial cross loadings increases with military rank. Within the highest status group the two domains can no longer be clearly distinguished. This is the same result found in the cognitive ability groups. Since none of the other personality dimensions showed substantial cross loadings on any of the other dimensions we conducted all further analyses with only these two dimensions: neuroticism and conscientiousness.

To sum up this first part, we believe that the high correlation between conscientiousness and neuroticism was caused by situational demand. This situational demand can be regarded as a systematic measurement error. This variance adds on to the trait variance and causes the correlations between the traits to increase.

Table 2.2. Exploratory factor analysis within military groups

<i>Military groups</i>						
recruits + students		non-commissioned officers		officers		
	NE	CO	NE	CO	NE	CO
N1	.78		.81		.41	-.66
N2	.74		.72	-.37	.41	-.67
N3	.79		.79		.31	-.72
N4	.74		.70	-.33	.46	-.61
N5	.37	-.61	.45	-.57		-.77
N6	.47	-.50	.46	-.59		-.75
C1	-.32	.60		.71		.78
C2		.65		.74	.37	.69
C3		.63		.72		.74
C4		.77		.78	.33	.67
C5		.70		.75		.77
C6		.70	-.34	.67		.77

Annotations. For facet names see figure 2.1. Loadings smaller than .30 have been omitted.

To test this hypothesis, we will reanalyze the data using structural equation modeling. We will conduct a multigroup analysis using two bimodal models (model 1a and 1b). Both models contain the facets of the personality dimensions as

manifest variables. Each facet represents the sum of the eight items measuring the facet. The bimodal models pay attention to the fact that the variables consist of two variance components: trait as well as state variance. The trait variance represents the personality of the participants. The state variance, however, stands for the situational demand the participants felt. Therefore, besides letting each manifest variable load on the trait factor it is supposed to measure, we also let each manifest variable load on a state factor. We call this factor *fake*. On the trait side, model 1a represents a Varimax rotated solution. This means there are no correlations between the personality dimensions. Model 1b additionally contains correlations between the personality dimensions and thus represents a correlated traits model.

Moreover, to shed some more light into the discussion whether the BIG 5 are correlated or not, we also tested two models *without* the latent variable *fake* (model 2a and 2b). In model 2a, there is no correlation between the personality dimensions. Thus, the model represents a Varimax rotated solution. Model 2b includes correlations between personality dimensions.

If the hypothesis that the situational demand is responsible for the higher correlations between personality dimensions was correct, the mean of the latent variable *fake* in models 1a and 1b should increase with the military rank. This would indicate that the situational pressure was felt stronger by higher military ranks. Moreover, the correlations between the personality dimensions in model 1b should be smaller compared to the correlations in model 2b.

### 2.3. Statistical Analyses

Confirmatory factor analyses (maximum likelihood) were conducted using AMOS 5.0. One important assumption for such an analysis is a multivariate normal distribution. We tested this assumption with the Mardia Test. The result of the Mardia Test (multivariate kurtosis = 25.987, *c.r.* = 11.163,  $p < .001$ ) shows a violation of the assumption. In such a case the  $\chi^2$ -test is too liberal and the p-value should be corrected. Therefore, we conducted a Bollen-Stine bootstrap with  $N = 200$  samples to correct the p-value of the  $\chi^2$ -test.

The assessment of the global-goodness-of-fit was based on the Standardized Root Mean Square Residual (*SRMR*) and the Root Mean Squared Error of Approximation (*RMSEA*) as recommended by Hu and Bentler (1999). The authors also give some advice regarding possible cutoffs for the indices. Thus, the *SRMR* should be lower or equal to .11 and the *RMSEA* should be less than .06 for  $N > 250$  and less than .08 for  $N < 250$ . Additionally, we looked at the Comparative Fit Index (*CFI*) as advised by Beauducel and Wittmann (2005). According to Hu and Bentler the *CFI* should have a value of approximately .95. Marsh, Hau and Wen (2004) criticized these "golden rules" and pointed out that the recommended cutoffs are very restrictive and hardly achievable when using personality questionnaires. Nevertheless, we will apply the cutoffs, keeping in mind that they are very strict when dealing with personality questionnaires.

In order to conduct the multigroup analysis and the latent mean comparisons, the intercept for each manifest variable had to be set equal in the three groups respectively (see Byrne, 2001). Moreover, all latent means were fixed at zero in the group containing recruits. Thus, the means in the other groups represent the

increase in the latent variable compared to the recruit group. These values can be interpreted as effect sizes since the latent means are  $z$  – scores.

To see which of the four models (1a, 1b, 2a, 2b) fits the data best, the fit indices should be compared. Using  $\chi^2$  - difference tests models 1a versus 1b and 2a versus 2b can be compared directly. However, since not all of the models are nested, a  $\chi^2$  - difference test cannot be applied to compare the models with and without the latent variable *fake*. Thus, an information criterion (AIC) is used to compare the models directly. The lower the criterion is, the better a model fits the data. The information criterion also considers model complexity. This means, if a model only fits better because of its higher complexity, the information criterion would be larger than for a less complex model.

## 2.4. Results

### 2.4.1. *Sample description*

A detailed sample description can be found in Toomela (2003). Table 2.1 displays means and standard deviations for the military groups in the cognitive ability test. Table 2.3 contains means and standard deviations for the personality dimensions of the three groups in our analysis.

As can be seen, the level of neuroticism falls from recruits to non-commissioned officers and the level of conscientiousness rises. Both changes were significant and the effect sizes were small. The same can be said for all other comparisons (except for agreeableness) between the recruits and the non-commissioned officers. However, the non – commissioned officers and the officers no longer differ significantly in any of the personality dimensions.

Table 2.3. Personality dimensions for the 3 military groups.

	<i>Military groups</i>									<i>Comparisons</i>		
	(1) recruits + students			(2) non-commissioned officers			(3) officers			(1) vs (2)	(1) vs (3)	(2) vs (3)
	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>d</i>	<i>d</i>	<i>d</i>
NE	421	85.1	24.8	201	75.8	27.6	248	74.2	25.8	-0.35***	-0.41***	-0.06
EX	421	111.7	22.6	201	116.7	25.2	248	117.8	22.3	0.21*	0.26**	0.05
OE	421	97.9	16.4	201	102.5	19.5	248	103.3	18.1	0.26**	0.30**	0.04
AG	421	108.2	14.4	201	111.3	18.2	248	110.9	15.2	0.19	0.17	-0.02
CO	421	115.8	21.7	201	120.7	24.2	248	122.5	21.6	0.22*	0.30**	0.08

Annotations. NE = Neuroticism; EX = Extraversion; OE = Openness to Experience; AG = Agreeableness; CO = Conscientiousness. The comparisons depict Cohen's *d* using the pooled standard deviation as denominator. Bonferroni adjusted significance levels are \* =  $p \leq .05$ ; \*\* =  $p \leq .01$ ; \*\*\* =  $p \leq .001$ .

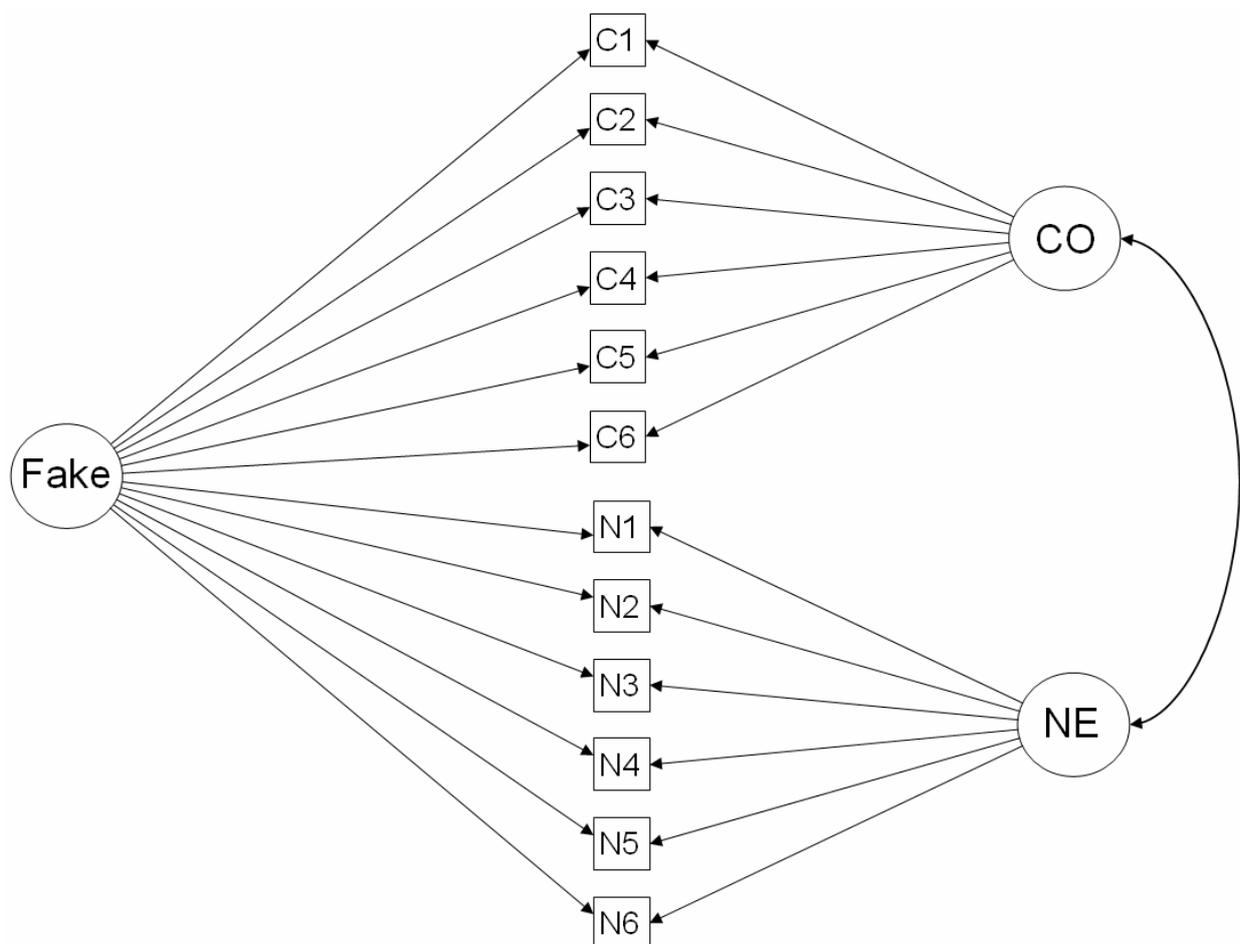
#### 2.4.2. Structural equation analysis

The following analysis combines procedures of multigroup analysis (Byrne, 2001) and *aspects* of Latent-State-Trait Theory (e.g. Steyer, et al., 1999). The observed variables in the models are the twelve personality facets from the NEO-PI-R (N1 to N6 for neuroticism and C1 to C6 for conscientiousness). All models contain two latent variables called NE for neuroticism and CO for conscientiousness which draw variance from the corresponding manifest variables. Moreover, the bimodal models include a third latent variable, called *fake* which draws variance from all observed variables. This variable represents the situational demand (state) and reflects the assumption that some part of the observed variance was due to the situation. Figure 2.1 illustrates the correlated bimodal model (1b) described above.

The first step in a multigroup approach should always be to test the model fit in one of the groups. Since the group of interest is the group of officers, we first tested the model in this group. Overall the model revealed an acceptable fit:  $\chi^2 [42] = 112.001$ , Bollen-Stine  $p$ -value = .005,  $SRMR = .038$ ,  $RMSEA = .082$  (90% confidence interval: .064 - .101),  $CFI = .97$ . Thus, the multigroup analysis was conducted next.

The results of the multigroup analyses with the four models can be found in table 2.4. Comparing the uncorrelated traits solutions with the correlated traits models shows that the correlated traits models (1b and 2b) reveal a better fit in both cases. The  $\chi^2$  - difference test between model 1a and 1b showed a significant result:  $\Delta\chi^2$  ( $\Delta df$ ) = 19.02 (3),  $p < .001$ . The same holds true for the comparison between model 2a and 2b:  $\Delta\chi^2$  ( $\Delta df$ ) = 385.15 (3),  $p < .001$ . Thus, considering the relationship between the personality dimensions always yields a significantly better result than an uncorrelated solution.

Figure 2.1. Structural Equation Model of model 1b.



Annotations. Facet names of personality domains: N1: anxiety; N2: angry hostility; N3: depression; N4: self-consciousness; N5: impulsiveness; N6: vulnerability; C1: competence; C2: order; C3: dutifulness; C4: achievement striving; C5: self-discipline; C6: deliberation. Error variances have been omitted.

Moreover, the models, which only contained the personality dimensions, but not the *fake* variable, revealed worse fits compared to the bimodal models. According to the *AIC*, the model with correlated personality dimensions and a latent situational demand variable represents the data best.

The latent correlations between neuroticism and conscientiousness in model 2b were  $r = -.62$ ,  $r = -.70$ , and  $r = -.75$  and controlling for situational demand (model 1b)  $r = .49$ ,  $r = .33$ , and  $r = .39$  for recruits plus students, non-commissioned officers, and commissioned officers respectively. Thus, the correlations in the bimodal model are clearly smaller, but most notably have reversed directions. However, the correlations are corrected for attenuation. In order to compare the correlations, which resulted when controlling situational demand, with the original bivariate correlation, we calculated the construct reliabilities for conscientiousness and neuroticism according to the equation by Hancock and Mueller (2001). The construct reliabilities for conscientiousness in model 1b were  $r_{tt} = 0.42$ ,  $r_{tt} = 0.24$ , and  $r_{tt} = 0.28$  for recruits plus students, non-commissioned, and commissioned officers respectively. The results for neuroticism in model 1b were  $r_{tt} = 0.66$ ,  $r_{tt} = 0.75$ , and  $r_{tt} = 0.66$  for recruits plus students, non-commissioned, and commissioned officers respectively.

Thus, the correlations between conscientiousness and neuroticism when controlling for situational demand were  $r = .21$ ,  $r = .14$ , and  $r = .17$  for recruits plus students, non-commissioned officers, and commissioned officers respectively.

Table 2.4. Model fit

	$\chi^2$ (df)	<i>p</i> - value	SRMR	RMSEA (90% CI)	CFI	AIC
Model 1a	483.30 (144)	.005	.049	.052 (.047 - .057)	.95	735.30
Model 1b	458.28 (141)	.005	.039	.051 (.046 - .056)	.95	716.28
Model 2a	1337.05 (182)	.005	.248	.086 (.081 - .090)	.81	1513.05
Model 2b	951.90 (179)	.005	.093	.071 (.066 - .075)	.87	1133.90

Annotations. SRMR = Standardized Root Mean Residual; RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; AIC = Aikake's Information Criterion.

Table 2.5 contains the loadings for the three groups in the 4 models. It can be seen that all manifest variables consist of variance due to trait, as well as to situational demand (state). However, there are changes in the loadings when situational demand is controlled (model 1a and 1b). Within the models 2a and 2b all manifest variables show high loadings on their respective personality dimensions. Considering the situational influence, the picture changes. Conscientiousness is dominated by the variable C2 (order), especially within the groups of non-commissioned and commissioned officers. The other conscientiousness facets have higher loadings on the situational factor. Only within the recruits group the loadings on both sides of the bimodal models are comparable. For neuroticism the pattern is different. The loadings are higher on the personality dimension or comparable to the loadings on the situational demand variable.

Table 2.6 includes the means and variances for all latent variables for the three military groups in the different models.

As mentioned earlier, the latent means for the recruits were fixed at zero. The hypothesis regarding the situational demand was that the latent mean for *fake* should increase with military rank. The results show that both officer groups differ significantly from the recruits group. The difference between the officer groups is

significant as well ( $t_{(447)} = 4.74, p < .001$ ). The effect sizes can be regarded as moderate to large.

Table 2.5. Loadings

	recruits+ students				non-commissioned officers				officers			
	Loadings on Conscientiousness											
	1a	1b	2a	2b	1a	1b	2a	2b	1a	1b	2a	2b
C1	0.08	-0.01	0.72***	0.75***	-0.15	-0.11	0.77***	0.79***	0.01	-0.15	0.80***	0.82***
C2	0.52 <sup>a</sup>	0.41 <sup>a</sup>	0.70 <sup>a</sup>	0.69 <sup>a</sup>	0.42 <sup>a</sup>	0.34 <sup>a</sup>	0.76 <sup>a</sup>	0.74 <sup>a</sup>	0.61 <sup>a</sup>	0.45 <sup>a</sup>	0.80 <sup>a</sup>	0.79 <sup>a</sup>
C3	0.42***	0.39***	0.69***	0.67***	0.34***	0.35***	0.73***	0.72***	0.27***	0.17*	0.80***	0.79***
C4	0.54***	0.45***	0.82***	0.80***	0.08	0.10	0.65***	0.63***	0.31***	0.18*	0.73***	0.73***
C5	0.36***	0.24**	0.84***	0.86***	0.22*	0.16*	0.85***	0.86***	0.32***	0.19*	0.87***	0.87***
C6	0.23***	0.15*	0.54***	0.54***	0.06	-0.02	0.54***	0.55***	0.22*	0.12	0.72***	0.72***
	Loadings on Neuroticism											
	1a	1b	2a	2b	1a	1b	2a	2b	1a	1b	2a	2b
N1	0.65 <sup>a</sup>	0.62 <sup>a</sup>	0.84 <sup>a</sup>	0.82 <sup>a</sup>	0.72 <sup>a</sup>	0.71 <sup>a</sup>	0.82 <sup>a</sup>	0.79 <sup>a</sup>	0.67 <sup>a</sup>	0.66 <sup>a</sup>	0.85 <sup>a</sup>	0.83 <sup>a</sup>
N2	0.48***	0.45***	0.71***	0.71***	0.53***	0.51***	0.80***	0.80***	0.40***	0.43***	0.77***	0.77***
N3	0.66***	0.62***	0.87***	0.86***	0.76***	0.75***	0.90***	0.87***	0.56***	0.56***	0.87***	0.87***
N4	0.56***	0.53***	0.76***	0.75***	0.50***	0.49***	0.76***	0.76***	0.46***	0.48***	0.72***	0.72***
N5	0.19***	0.13	0.42***	0.44***	0.18*	0.15*	0.53***	0.55***	0.16*	0.15*	0.59***	0.61***
N6	0.18**	0.16*	0.68***	0.72***	0.26***	0.24***	0.75***	0.78***	0.31***	0.33***	0.77***	0.80***
	Loadings on Fake											
	1a		1b		1a		1b		1a		1b	
C1	0.82 <sup>a</sup>		0.81 <sup>a</sup>		0.90 <sup>a</sup>		0.86 <sup>a</sup>		0.86 <sup>a</sup>		0.86 <sup>a</sup>	
C2	0.51***		0.59***		0.66***		0.68***		0.66***		0.75***	
C3	0.55***		0.59***		0.65***		0.65***		0.74***		0.78***	
C4	0.63***		0.70***		0.63***		0.62***		0.66***		0.70***	
C5	0.76***		0.81***		0.81***		0.83***		0.80***		0.84***	
C6	0.48***		0.52***		0.56***		0.59***		0.69***		0.71***	
N1	-0.53***		-0.57***		-0.44***		-0.46***		-0.57***		-0.57***	
N2	-0.51***		-0.54***		-0.58***		-0.60***		-0.66***		-0.64***	
N3	-0.57***		-0.61***		-0.53***		-0.55***		-0.67***		-0.67***	
N4	-0.51***		-0.54***		-0.56***		-0.58***		-0.55***		-0.54***	
N5	-0.41***		-0.44***		-0.53***		-0.55***		-0.63***		-0.64***	
N6	-0.83***		-0.81***		-0.79***		-0.80***		-0.77***		-0.75***	

Annotations. For facet names see figure 2.1. <sup>a</sup> = path had starting value of 1, therefore no significance test possible. \* =  $p \leq .05$ ; \*\* =  $p \leq .01$ ; \*\*\* =  $p \leq .001$ .

The results for conscientiousness are interesting. Not paying attention to the situational demand, conscientiousness increases significantly with military rank

( $t_{(447)} = 6.43$ ,  $p < .001$  for 2a and 2b). Controlling for situational influence, the mean for conscientiousness was significantly larger for the non-commissioned officers than for the recruits. However, the commissioned officers had a significantly smaller mean than the non-commissioned officers ( $t_{(447)} = 3.58$ ,  $p < .001$  for 1a and 1b). Moreover, the commissioned officers did not differ significantly from the recruits in their conscientiousness. For the factor neuroticism the mean decreases significantly with military rank in model 2a and 2b ( $t_{(447)} = 15.49$ ,  $p < .001$  for non-commissioned vs. commissioned officers). Controlling for situational influence, neuroticism also falls significantly with military rank ( $t_{(447)} = 9.38$ ,  $p < .001$  for non-commissioned vs. commissioned officers). Regardless of the model all latent variables except one have significant variances. The exception is conscientiousness in model 1b within the group of non-commissioned officers. That means controlling for situational influence and letting the personality dimensions correlate, conscientiousness does not have a variance significantly different from zero in the group of non-commissioned officers.

Table 2.6 Latent means and significance of comparison with recruits.

	recruits+students				non-commissioned officers				officers			
	1a	1b	2a	2b	1a	1b	2a	2b	1a	1b	2a	2b
CO	M	0	0	0	1.1	1.10**	3.58***	3.50***	.76	0.76	4.19***	4.11***
	S <sup>2</sup>	4.55***	2.74*	8.27***	7.90***	3.29*	2.14	11.51***	11.21***	8.21***	4.46*	13.74***
NE	M	0	0	0	-1.56*	-1.56***	-3.82***	-3.80***	-2.48***	-2.48***	-5.29***	-5.26***
	S <sup>2</sup>	11.09***	10.03***	18.17***	17.57***	13.03***	12.56***	16.03***	15.10***	8.87***	8.69***	14.38***
Fake	M	0	0		3.35***	3.35***			3.70***	3.70***		
	S <sup>2</sup>	10.04***	11.99***						11.30***	13.48***		

Annotations. CO = Conscientiousness, NE = Neuroticism, \* =  $p \leq .05$ ; \*\* =  $p \leq .01$ ; \*\*\* =  $p \leq .001$ . The mean in the groups of recruits and students was fixed at 0.

One could argue now that intelligence and not military rank is the key variable responsible for the increasing mean of the situational demand variable. In order to test this we included the variable cognitive ability group into the model and let it correlate with *fake*. This variable stems from the original data set and comprises the five different ability groups Toomela created. The correlations were  $r = .16$  ( $p < .05$ ) for the recruits,  $r = .10$  (n.s.) for the non-commissioned officers and  $r = -.15$  ( $p < .05$ ) for the commissioned officers. The structure of the model itself did not change. Thus, the influence of intelligence on the situational demand variable is either insignificant or very small.

## 2.5. Discussion

The goal of this paper was to reanalyze the data reported by Toomela controlling for situational influence. Toomela could not confirm the BIG 5 model in a group of highly intelligent men. In this group, the factors conscientiousness and neuroticism could not be differentiated. In the course of this reanalysis it was hypothesized that situational demand was the reason for this problem. The hypothesis was that not intelligence but military rank was the attribute responsible for the influence of situational demand. In order to prove this hypothesis, a multigroup structural equation model including aspects of LST was used to analyze the data again. The results confirm the hypothesis and show that the mean for the latent variable *fake* increases with military rank. This was still the case when cognitive ability was controlled. Moreover, the correlation between conscientiousness and neuroticism was much smaller when situational demand was controlled. In the course of this reanalysis two differing views of the BIG 5 were contrasted: an uncorrelated and a correlated model. The results show that a correlated model represents the data

better than an uncorrelated model. Moreover, controlling for situational demand further improves the model. All in all, the model with correlated personality dimensions and a situational demand factor fits the data best. It should be noted, however, that the direction of the relationship between the personality dimensions switched from negative to positive when controlling for situational demand.

### *2.5.1. Situational demand*

The results provided here show that situational demand (state) is a phenomenon which is felt differently by different groups of people. More clearly, in the data by Toomela it looks as if men with a higher military rank felt a greater need to obey the situational demand and thus appear as more prototypical military. The result was that they presented themselves as less anxious, more stable, hard-working and conscientious. However, a crucial amount of variance in the personality questionnaires was due to the situational demand. It has to be noted, that military rank and intelligence are confounded. In other words, with the present data it is not possible to differentiate between the amount of state variance due to military rank and due to intelligence. Thus, further studies should be conducted to separate these sources of variances.

Interestingly, the loadings on the situational factor are almost always higher than on the trait factors in all three groups and in both models (1a and 1b). Thus, situational demand seems to play an important role. To estimate the importance, a look at the variances of the personality dimensions is very informative. Not controlling for situational influence (models 2a and 2b), the personality dimensions both have large amounts of variance in all three groups. This changed when situational demand was controlled. Now, conscientiousness loses a big amount of

variance to the situational demand variable. That means a lot of the differences in conscientiousness were actually not due to different values in persons' personalities, but to the way they responded to the situational demand. This is especially evident in the group of non-commissioned officers. Here conscientiousness no longer has a significant amount of variance. In the group of officers conscientiousness does not disappear, but is dominated by a single facet: order. All other facets contain mostly situational variance. Thus, situational demand has two sides. First, it is felt differently by different groups. Recruits probably had the least ambition to appear as a military person. However, with increasing rank the demand was felt stronger. Secondly, the effect of the situational demand also is different in the groups. One could assume the more the situational demand was felt, the less personality variance should be left. However, this is not the case. The personality dimensions are not equally affected. This is in line with results by Pauls and Crost (2005b) who demonstrated the selective effects of faking, a form of situational demand. While neuroticism keeps a big amount of variance, conscientiousness seems to fall apart. Thus, conscientiousness seemed to have higher importance for the non-commissioned officers and officers in order to appear as a military person than neuroticism. At this point a note of caution is appropriate. The design of the reanalysis only contains aspects of a real LST design. Thus, the amount of variance due to situational demand is probably overestimated. After all, Deinzer et al. (1995) could show that it hardly exceeds 20%. Only experimental designs can bring more information into this matter. Summing up, there is evidence that a specific "faking" took place which caused the two personality dimensions neuroticism and conscientiousness to correlate highly. By splitting the variance into trait and situational demand variance the personality dimensions can be better distinguished.

### 2.5.2. *BIG 5 Models*

Another aspect of this reanalysis was to compare an uncorrelated and a correlated traits model and to find out which impact situational demand has on both. The results show that the correlated traits model fits better than the uncorrelated and that controlling situational demand also improves the model. The first result seems to indicate that it is inappropriate to regard the BIG 5 as uncorrelated. If one only looked at models 2a and 2b, this would clearly be the case. Here, the correlated model fits significantly better and the correlation found is in line with Mount et al. who reported a true correlation of .52 with a 90% confidence interval of .27 to .73. Moreover, the fit indices for the uncorrelated model (2a) are clearly worse than for the correlated model (2b). Thus, it seems as if an uncorrelated solution can be discarded. However, we would not draw this conclusion. In fact, on the basis of the results reported here we would advocate the uncorrelated traits solution. There are three reasons for this. First of all, controlling for situational demand decreases the correlations between personality dimensions drastically. Secondly, if the correction for attenuation was reversed, the correlations were only small or negligible, depending on the group. Of course, a correction for attenuation is not inappropriate. However, if reliabilities are as low as found in our study when controlling for situational demand, the correction for attenuation leads to highly increased estimations of the correlations. And after all, the original correlation we looked at was a bivariate correlation without a correction for attenuation, thus it is only fair to reverse the correction. And finally, the fit indices did not improve drastically from model 1a to 1b. One could argue that the  $\chi^2$  - difference test was significant. However, with large samples the power for this test increases and therefore, even small model violations would yield significant results.

Only regarding the fit indices, we draw the following conclusion: If situational demand (state variance) is controlled, an uncorrelated traits solution is not worse than a correlated traits solution. This is not a trivial interpretation. Costa and McCrae (1992b) attributed the missing orthogonality of the BIG 5 scales to a lack in discriminant validity. This however must not necessarily be true. The results reported here indicate that not a lack in discriminant validity but the common influence of situational demand might lead to violations of orthogonality. To our knowledge, this aspect has received very little attention in research and practice. Therefore, we hope to inspire more research on this topic.

### *2.5.3. Relationship between personality dimensions*

Interestingly, the correlation between neuroticism and conscientiousness reversed direction, when controlling situational demand. Mount et al. recoded neuroticism so that all personality dimensions have positive intercorrelations. Thus, the correlation found controlling for situational demand is in contrast to the meta-analytical findings. An obvious reason for this is that the meta-analysis did not consider situational influence which always has an impact (Deinzer, et al., 1995). The present results would mean that taking the variance caused by situational demand out of the trait variance, conscientious people tend to be anxious as well. One has to keep in mind however, that conscientiousness no longer contains an equal amount of variance from all facets but is dominated by order. Then the positive correlation is not counterintuitive. If one is rather anxious, it will be beneficial to be more orderly to calm the anxiety. This is, however, just an interpretation and needs to be examined further.

#### *2.5.4. Latent personality means*

It is also very informative to look at the latent personality means in the three groups when controlling situational influence. Toomela's original analysis yielded results which indicated that neuroticism falls with increasing intelligence while conscientiousness increases. These results are hard to interpret since both dimensions could not really be distinguished and the facets showed substantial cross-loadings. Nevertheless, even when controlling for situational demand the mean in neuroticism falls with military rank and thus with increasing intelligence. However, the mean for conscientiousness first increases (non-commissioned officers) and then decreases again in the group of officers. In other words, up to a certain extent intelligence and conscientiousness have a positive relationship which switches direction beyond this point. This result is in line with the findings by Moutafi et al. (2003) (see also LaHuis, Martin & Avis, 2005) who could confirm a negative relationship between conscientiousness and intelligence. Their conclusion was that a low ability can be compensated by a conscientious work attitude while a high ability does not need a conscientious work style.

#### *2.5.5. Limitations and further research*

One major drawback of this paper is the fact that this is a reanalysis and not an experimental study. We can only hypothesize that situational demand led higher military ranks to portray stronger faking behavior. One could argue that there are numerous other possible explanations for the differences in the situational demand factor. For example age, amount of life experience, kind of personality, or intelligence. At least for intelligence the analysis conducted here show that it has

little or no influence on the situational demand variable. The fact that the influence is positive for recruits and negative for commissioned officers actually strengthens the hypothesis. Intelligent recruits felt the situational demand and faked at least a little, maybe to please the test conductors. Commissioned officers on the other hand did cheat less with increasing intelligence. One could hypothesize this was the case because they did not want to cheat too obvious. Of course, these are mere speculations and most of all the influence was rather very small. Concerning the alternative explanation that age was responsible for the effects found, a recent meta-analysis by Roberts, Walton, and Viechtbauer (2006) showed that changes in personality over the course of several years can only be regarded as modest. Nevertheless, the effect sizes found in the data by Toomela between the different military groups were more than modest if situational demand was not controlled. Thus, it is reasonable that differences in age alone cannot explain the effects found. However, only with the help of an experimental design would it be possible to find clearer evidence for this hypothesis. Such a design should use Latent – State – Trait – Theory to determine the amount of situational impact and also pay attention to the reason why situational demand is felt. Only this way a confoundation with intelligence and other variables as in the present reanalysis can be avoided. It would then also be possible to shed light onto the ability which enables people to fake a questionnaire. Clearly, it must be an ability. As Pauls and Crost could show, not all dimensions are faked but only those believed to be relevant. Judging relevance must be connected to an ability. This raises the question of the role of intelligence. Another step should be to test situational demand in real life situations. A selection scenario seems very appealing in this regard. In such a context it would be interesting to find out whether the predictive power of personality questionnaires

comes from the personality variance they contain or from the ability to fill out the questionnaire in line with the situational demand.

Summing up, there is evidence to assume that situational demand was responsible for the large, negative correlation between conscientiousness and neuroticism. Thus, situational demand has an influence on the structure of the BIG 5 and should be regarded with more care as hitherto since by doing this the proposed uncorrelated personality structure is more likely to appear. All in all, regarding the current results an uncorrelated traits model seems very plausible if situational influence is controlled.

**Chapter 3: Personality Structure: A Question of Trait and State  
Interaction**

## Chapter 3. Personality Structure: A Question of Trait & State Interaction

Personality has been the interest of many psychologists and researchers worldwide. Especially in the last twenty years models derived from the psycholexical branch of research have become popular (e.g. Costa & McCrae, 1992b; Goldberg, 1981). Most notably, the five factor model (FFM) has enjoyed a lot of attention (Aluja, Garcia, Garcia & Seisdedos, 2005; Barrick, Mount & Judge, 2001; Chamorro Premuzic & Furnham, 2004; Digman, 1997; Mount, et al., 2005). Goldberg (1992) even wrote : "Gradually, agreement has been growing about the number of orthogonal factors needed to account for the interrelations among English-language trait descriptors" (p. 26). However, the FFM has been challenged on several grounds, most prominently: number of factors, methodological approach, orthogonality, and impact of social desirable responding. All of these points address the question of construct validity. The present study aimed at this question, but also investigates problems regarding criterion validity.

### 3.1. Critic on the construct validity of the FFM

The number of factors has been debated for a long time already (Costa & McCrae, 1992a; Eysenck, 1992). Even though the debate was heated, the question does not seem to be solved as new models with more or less than five factors emerge (Ashton, et al., 2004; Saucier, 2003; Saucier, et al., 2005). However, there are also questions regarding the methodological approach. Probably one of the most prominent adversaries of the FFM and specifically factor analysis as the main

tool used is Jack Block (1995). Block criticizes that factor analysis always is bound to rely on a number of subjective decisions. These decisions influence the number of factors extracted as well as the relationship between those factors. Costa and McCrae (1995) replied to the critique stating that there is mounting evidence for the so-called five-factor hypothesis and that it is a "serviceable model". The third major concern is the question of the orthogonality of the factors. Mount et al. (2005) conducted a meta-analysis and found substantial correlations between personality factors which range up to a true correlation of .50 between neuroticism and conscientiousness. Costa and McCrae attribute this to a lacking discriminant validity and advice users of their questionnaire to compute factor scores with the factor loadings derived from the American normative sample to avoid correlations between personality factors. Biesanz and West (2004) claim that an evaluative bias causes correlations between the personality factors.

Recently Ziegler, Bühner, and Toomela (submitted-b) showed that correlations between personality factors might be due to situational demand. They argued that situational demand leads to an increased state variance which adds on to the factors and, like correlated errors, causes correlations to rise. They based their line of argumentation on the Latent-State-Trait model (LST) by Steyer (e.g., Steyer, et al., 1992; Steyer, et al., 1999). In LST models personality questionnaire variance can be split into trait, state, interaction between trait and state and error. Using this model Deinzer and colleagues could show that up to twenty percent of personality questionnaire variance is due to state or interaction between state and trait (Deinzer, et al., 1995). This impact of situational demand (faking) on the structure of the FFM has been ruled out elsewhere. Smith, Hanges, and Dickson (2001) analyzed a student as well as an applicant sample and found the FFM to be adequate in both samples. However, they allowed for correlations between personality factors, which

is not in line with the theoretical model. Moreover, looking at the correlations shows that they are higher within the applicant sample. Smith and Ellingson (2002) used the same method as Ziegler, Bühner, and Toomela to separate trait variance and what they called a method factor (social desirability). They compared this model in a student and an applicant sample and found no different loading patterns between both groups. They interpreted this as prove that construct validity is not impacted by social desirability (situational demand). This interpretation can be challenged on two grounds. First, as Ziegler, Bühner, and Toomela could show, faking can occur even in a sample where no one would suspect faking. Thus, the indifferent loading patterns might be due to a situational demand felt by the students as well. In line with this argument are findings by Zickar, Gibby, and Robie (2004) who reported that a large number of incumbents without any need for faking still faked in a personality questionnaire. Secondly, Smith and Ellingson only compared loading patterns and not latent means. The latent means would indicate whether the applicants actually faked more than the students. A flaw of both studies is that only one measurement point was used. Thus, a clear separation of trait and state was not possible.

A research line which is closely related to the state trait question is concerned with the impact of social desirability or faking on construct validity (e.g., Collins & Gleaves, 1998; Ellingson, et al., 2001; e.g., Stark, et al., 2001; Zickar & Robie, 1999a). Basically results indicate that faking causes increased correlations between personality factors. This aspect automatically raises concerns regarding construct and criterion validity. These concerns will be discussed next before we turn to the question what exactly social desirability or faking might be.

### 3.2. Critic on the criterion validity of the Big 5

The question of construct validity also impacts criterion validity. The size of the correlations between personality factors has a direct impact on the criterion validity of each factor. Usually, within validity studies criteria are regressed on all five factors simultaneously (Chamorro Premuzic & Furnham, 2003; Furnham & Chamorro-Premuzic, 2004; Furnham, et al., 2002; Salgado, 2003). However, if predictors are correlated, the amount of variance one predictor explains exclusively decreases, limiting chances of finding significant regression weights. One could argue that this is no problem since the overall amount of explained variance is not affected. This argument only holds true as long as one is not interested in finding specific predictors. Thus, non-orthogonality affects criterion validity.

There is evidence that the correlations between personality factors increase in applicant settings (e.g., Collins & Gleaves, 1998; Ellingson, et al., 2001; e.g., Pauls & Crost, 2005b; Schmit & Ryan, 1993; Zickar & Robie, 1999a). Schmit and Ryan coined the term "ideal employee factor" since they found a sixth factor on which all facets had loadings. Pauls and Crost could show that not all personality dimensions are affected by faking instructions but only those believed to be job-relevant. A similar result was obtained by Furnham (1997) who found an impact of faking only for three of the five factors, namely conscientiousness, agreeableness and neuroticism. However, even though there is evidence that personality measures can be faked and that this faking affects the construct validity, there is also evidence that the criterion validity of such measures remains untouched (e.g., Salgado, 2003). Taking these results into account one could ask the question where does the criterion validity of personality measures stem from? Is it really personality that predicts performance or is it the additional information which causes correlations

between factors to rise? If the latter is true, what exactly is this additional information which results from faking?

Connected to this is the fact that faking increases mean scores in personality questionnaires (e.g., Viswesvaran & Ones, 1999; Ziegler, et al., submitted-a). This topic has been and still is a major concern for many practitioners keeping them from using personality questionnaires. The reason for this is that some practitioners doubt that faked personality measures still assess personality. To rule out this argument, the impact of social desirability on criterion validity has been investigated. Ones, Viswesvaran and Reiss (1996) conducted a meta – analysis and stated that social desirable responding does not impact criterion validity, but is in fact due to actual individual differences in neuroticism and conscientiousness. Paulhus (2002) gives an excellent overview of the emergence of social desirability as a construct and also claims that it represents a trait. Research on social desirability has one important flaw, though. It is assumed that questionnaires constructed to measure social desirable responding cannot be faked themselves. This, however, seems to be a wrong presumption as Pauls and Crost (2004) could show. Thus, within the present study no social desirability measures will be applied to isolate faking. Instead a more sophisticated methodological approach, namely a LST model in combination with an Individual Causal Effects (ICE) model (Steyer, 2005), will be chosen.

### 3.3. Substance vs. style

As some of the above mentioned research indicates, there have been attempts to find out what exactly faking in personality questionnaires is. Some researchers believe it is social desirability (Paulhus, et al., 2003), fuelling discussions whether social desirability impacts criterion validity or construct validity or both (Ellingson, et al., 2001; Ones, et al., 1996; Smith & Ellingson, 2002; Smith, et al., 2001). Other researchers claimed that faking is an ability (Christiansen, Burns & Montgomery, 2005; Nguyen, Biderman & McDaniel, 2005; Pauls & Crost, 2005a; Peeters & Lievens, 2005; Vasilopoulos, Cucina & McElreath, 2005) and tried to determine its correlation with measures of cognitive ability.

However, as mentioned above, the first line of research is problematic since measures of social desirability can be faked as well. The second line of research is impaired by a different problem. Usually, faking is operationalized as the difference in personality mean scores between an honest and a faking condition. Difference scores, however, have low reliabilities, thus the found correlations between faking and cognitive ability which were mostly small to moderate might be an underestimation. Pauls and Crost (2005b) residualized the honest variance out of the faked variance and used those residuals as measures of faking in order to circumvent the problems associated with difference scores. They found a relation between general intelligence and the amount of faking. Moreover, they looked at the relationship between faking and self-deceptive enhancement and impression management (BIDR) as well as measures of self efficacy of positive self-presentation beliefs. Their results indicated that the latter was a better predictor of faking than both of Paulhus measures. All in all, there is growing evidence that faking can be

considered as a trait, yet a clear answer regarding the components has not been found.

#### 3.4. Present study

Summing up, there is evidence for state variance within trait measures of personality (Deinzer et al.). Moreover, an increase in situational demand and thus, in state variance increases correlations between personality factors. This was hypothesized by Ziegler, Bühner, and Toomela. Smith and Ellingson (2002) came to the conclusion that situational demand does not impact construct validity. However, both studies did not apply a complete LST model with at least two measurement points which would allow for a separation of state and trait variance. To overcome this limitation, personality measures were conducted twice within the present study. Moreover, the present study consisted of two groups, one of which received a faking instruction at the second time while the other group was asked to answer honestly at both times. Undoubtedly, the situational demand in the first group will be higher due to the faking instruction. Thus, the question of construct validity is tapped.

Besides the possible impact of situational demand on construct validity the impact on criterion validity is also investigated: Where does the criterion validity of personality questionnaires originate – trait variance or state variance? In order to investigate this question the present study used an Individual Causal Effects (ICE) model (Steyer, 2005) in combination with a LST model. ICE models basically are control group designs using structural equation modeling and two measurement points. As mentioned above, within the present study two groups (control and experimental group) worked on a personality questionnaire without special instructions. After an interval they completed the same questionnaire again. This

time, however, the experimental group was given a faking instruction. Since the control group did not receive a special treatment, any changes within the experimental group which do not appear in the control group can be causally attributed to the treatment. Through combination with an LST model it will also be possible to split variance into trait and state variance and investigate their predictive power separately. Finally, the character of the state variable will be investigated. Is it an ability or is it a personality trait?

To sum up, the present study had the aim of investigating the impact of situational demand or state on construct and criterion validity of the FFM. The following hypotheses will be tested: (1) Situational demand impacts measures of personality structure and causes correlations between personality factors. Controlling for situational demand will render uncorrelated personality factors. (2) Situational demand increases within a specific applicant setting compared to an average laboratory situation. (3) Furthermore, the question of criterion validity of trait and state variance will be explored: Is it state or trait variance which causes the predictive validity of personality questionnaires? (4) Finally, the character of the state variable will be explored: Is it an ability or is it a personality trait?

### 3.5. Methods

#### 3.5.1. *Participants and procedure*

The total sample consisted of 186 (140 women and 46 men) psychology students enrolled at the Ludwig-Maximilians-University in Munich. The average age and semester were 23.03 ( $SD = 5.77$ ) and 1.89 ( $SD = 2.01$ ), respectively.

As described above, the experiment was conducted with two different groups, a control group (CG,  $N = 94$ , 27 men) and an experimental group (EG,  $N = 92$ , 19 men). All tests were administered with a computer in a laboratory containing seven work stations. Group assignment was random. When participants arrived at the laboratory, they first had to fill out a personality questionnaire following the normal instructions. Afterwards participants in both groups worked on a cognitive ability test, a lexical knowledge test and a measure of self efficacy beliefs of positive self-presentation, which lasted for about two hours. After those tests participants received a specific instruction depending on their group assignment. In the CG participants were told that they should not be surprised if they knew the upcoming test and should answer the questions as honestly as possible. Within the EG, however, participants received a fake good instruction. Rogers (1997) stated a list of advisable design features in a faking experiment. Most importantly, the sample should have practical relevance and a realistic scenario accompanied by a note of caution regarding obvious faking should be given. Within the present study psychology students from early semesters were used as a sample. The practical relevance for this sample is high, since universities have to select students for admission. A personality test might be used by some universities as selection tool. The scenario was in line with this practical point and gave the experiment

ecological validity. Participants in the faking group were given the following special instruction:

Universities have to select their students. For this task a number of instruments like the following are being tested right now. Please imagine that you are participating in a student selection procedure. Of course, it is your goal to get an admission as a psychology student. Therefore, you have to fill out the following questionnaire in a way that assures your admission. However, you have to be careful since a test expert will check the results for obvious faking and you do not want to be spotted.

After reading this specific instruction, all participants filled out the personality questionnaire for a second time. Upon completion, they were asked whether they had followed the specific instructions or not. All participants stated that they had followed it. Finally, they were thanked and dismissed.

### *3.5.2. Test materials*

Both groups had to fill out the NEO – PI – R (Costa & McCrae, 1992b) twice. This well-known personality questionnaire consists of 240 items which assess six facets for each of the five factors of the FFM, i.e., neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. Thus, each facet contains eight items. Items ask participants to rate themselves in typical behaviors or reactions on a five-point Likert scale, ranging from “strongly disagree” to “strongly agree”. Alphas for the factors range from .87 to .92. For the facets they range from .53 to .85.

The cognitive ability test was the basic module from the Intelligence Structure Test (IST 2000 R) (Amthauer, Brocke, Liepmann & Beauducel, 2001). This module consists of nine subtests with twenty items each. Three subtests combine to measure verbal intelligence, numerical intelligence, and figural intelligence, respectively. Those three kinds of intelligence are combined to a reasoning score (Cronbach  $\alpha = .97$ ). The theoretical basis for the reasoning measure ( $g_f$ ) has been shown (Beauducel, Brocke & Liepmann, 2001).

As a measure of crystallized intelligence in the form of lexical knowledge, the Lexical Knowledge Test (Lexikon-Wissen Test, LEWITE, Wagner-Menghin, 1998) was used. The LEWITE is an adaptive test measuring crystallized intelligence ( $g_c$ ) in form of Rasch parameters (Cronbach  $\alpha = .86$ ). Participants have to complete sentences which define difficult words. Participants also have to state first whether they know the word or not, which allows the computation of a self-over-estimation score (SOE) reflecting how much participants overestimate their own ability. This measure has the same reliability.

Furthermore, the same questionnaire used by Pauls and Crost (2005b), the Efficacy of positive Self-Presentation Questionnaire (Mielke, 1990), was applied to get a measure of how much participants believe to be able to make others like them and think they are smart and capable people. All in all, 33 items have to be rated on a four point Likert scale ranging from "totally disagree" to "totally agree". The questionnaire has been described in more detail by Pauls and Crost. As those authors suggested, a total score for self efficacy beliefs of positive self-presentation (SEB) was computed. Cronbach  $\alpha$  (on item level) of this measure was .86. Within the present sample the total score correlated  $r = -.46$  ( $p < .01$ ) with neuroticism.

As criterion variable served the score from a statistics examination<sup>2</sup> which all psychology students have to take after their first semester. The variable was chosen because usually especially statistics is the most problematic subject and hence should have sufficient variance for prediction purposes. Besides this, the grade has been used in other studies as well (Furnham & Chamorro-Premuzic, 2004) and proved to be predictable by personality measures. The statistic examination consisted of 33 multiple choice items with an internal consistency of  $\alpha = .80$ . All participants took the examination at least 2 months after being subjected to the other tests used in the study.

### 3.5.3. Models

As described above, the study combines an ICE design and a LST design. Both methods make use of structural equation modeling. Thus, the analyses consisted of three different structural equation models: (1) an uncorrelated trait model (see figure 3.1), (2) a correlated trait model, and (3) a prediction model. Since it is assumed that not all personality factors are faked, the models only contain those factors which were faked (Pauls & Crost, 2005b). The experimental design allowed to split the variance into trait and state variance. Since the items were the same at both measurement times, correlated errors between identical facets were included representing method variance. It was hypothesized above that (1) situational demand impacts measures of personality structure and causes correlations between personality factors. Controlling for situational demand will render uncorrelated personality factors. If this holds true, bivariate correlations between personality

---

<sup>2</sup> 24 % of the participants did not take part in the statistic examination. Their score was estimated using SPSS 14.0 and the EM method. In order to get better estimates, scores from the tests taken between the personality questionnaires were also used for estimation

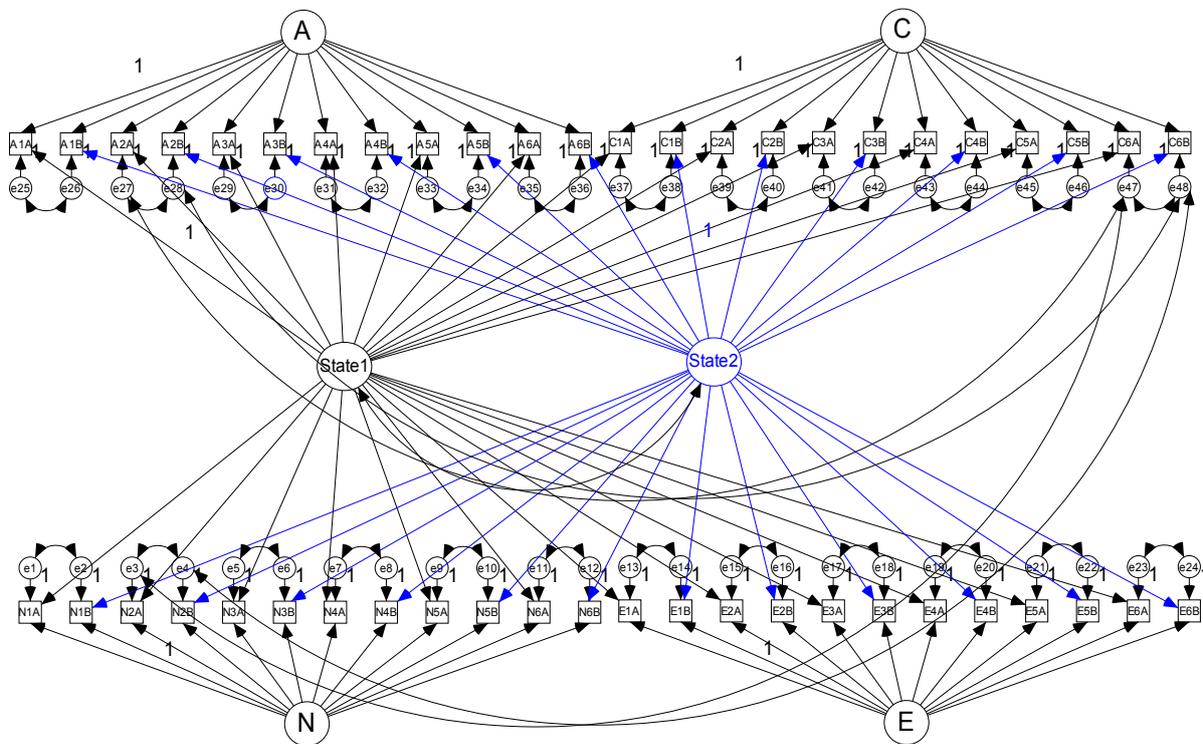
factors should be higher in the EG than in the CG if situational demand is not controlled and comparable as well as smaller if it is controlled. Moreover, model 1 should have a better fit than model 2. (2) It was proposed that situational demand increases within a specific applicant setting compared to an average laboratory situation. If this holds true, the latent mean for state 2 should be significantly higher in the EG than in the CG. Moreover, since the circumstances for the CG did not change, the correlation between state 1 and 2 should be high in the CG while it should be low in the EG due to the treatment, and, state 1 should be equal in both groups. (3) Furthermore, the question of criterion validity of trait and state variance will be explored. In order to do this, the points achieved in the statistics examination will be taken into the model and regressed on the personality factors as well as state 2. (4) Finally, the character of the state 2 variable will be explored. Correlations between state 2 and reasoning, lexical knowledge, self efficacy beliefs of positive self-presentation, and self-overestimation will be computed. Since all these measures were assessed at the same time, no regression was conducted.

#### *3.5.4. Statistical Analysis*

Besides the structural equation models, the data will be subjected to an analysis without controlling for situational demand using SPSS 14.0. Therefore, in a first step bivariate correlations between personality factors without control for situational demand will be compared for the two groups. In a second step scores on the personality factors will be compared between measurement points and between groups. Therefore, a multivariate analysis of variance with the independent variables CG vs. EG and time 1 (t1) vs. time 2 (t2) and the personality factors as

dependent variables will be conducted. If the faking treatment is successful, the interaction between group and repeated measurement factor should be significant.

Figure 3.1 Structure Equation Model



Legend: Facet names of personality domains: N1: anxiety; N2: angry hostility; N3: depression; N4: self-consciousness; N5: impulsiveness; N6: vulnerability; E1: warmth; E2: gregariousness; E3: assertiveness; E4: activity; E5: excitement-seeking; E6: positive emotion; A1: trust; A2: straightforwardness; A3: altruism; A4: compliance; A5: modesty; A6: tender-mindedness; C1: competence; C2: order; C3: dutifulness; C4: achievement striving; C5: self-discipline; C6: deliberation. The letter A behind the facet name refers to time 1 and B to time 2. Correlated errors between identical items represent method variance.

Both main effects are not insightful. The main effect group would tell whether the two groups differ at all and the main effect time would tell whether there was a change from time 1 to 2 at all. Of interest for the present study, however, are changes within groups across time and between groups at a given time. Therefore, results for the main effects will be omitted and t-tests will be conducted in a next step. Significance levels will be Bonferroni corrected for the number of significance tests conducted within each method. One-tailed tests will be conducted for the

comparison of CG and EG at time 2 since it can be expected that participants portrayed themselves as less neurotic, more extraverted, open minded, agreeable, and conscientious when asked to fake. For the comparison within the EG between both measurement points one-tailed tests will be applied as well for the same reason. For all other comparisons no hypothesis can be made since changes in both directions might occur. Thus, two-tailed tests will be applied.

Confirmatory factor analyses (maximum likelihood) were conducted using AMOS 6.0. Multivariate normal distribution was tested with the Mardia Test (multivariate kurtosis = 45.58, *c.r.* = 2.54,  $p < .05$ ) which showed a significant deviation. Therefore, Bollen-Stine bootstraps with  $N = 200$  samples were performed to correct the  $p$ -value for the  $\chi^2$  - tests. Moreover, the sample used in the present analysis is rather small compared with the number of parameters which need to be estimated. This disparity can lead to severe estimation problems. To avoid such problems and the possible mistakes in the interpretation the trait loadings in the CG were set equal for each trait. Results showing comparable trait loadings support this procedure (e.g., Allik & McCrae, 2004).

The assessment of the global-goodness-of-fit was based on the recommendations by Hu and Bentler (1999) as well as Beauducel and Wittmann (2005). Thus, the Standardized Root Mean Square Residual (*SRMR*, should be lower or equal to .11), the Root Mean Squared Error of Approximation (*RMSEA*, should be less than .06 for  $N > 250$  and less than .08 for  $N < 250$ ) and the Comparative Fit Index (*CFI*, should have a value of approximately .95) were used.

In order to conduct the multigroup analysis and the latent mean comparisons the intercept for each manifest variable had to be set equal in both groups (see Byrne, 2001). All latent means were fixed at zero in the CG (except for state 2). The latent

means in the EG can thus be regarded as effect sizes for the difference between both groups.

Comparison between the correlated and the uncorrelated model will be done by using a  $\chi^2$  - difference test as well as a comparison of the CFI (Cheung & Rensvold, 2001). Significant results are indicated by a difference in CFI larger than .01.

### 3.6. Results

Descriptive statistics for the used measures for both groups can be found in table 3.2. It can be seen that differences between CG and EG at time 1 were small and statistically insignificant.

#### *3.6.1. Without controlling situational demand*

Before looking at the results derived from structural equation modeling, the data were checked for effects which appear when situational demand is not controlled. The first step was a comparison of the bivariate correlations between personality factors at both measurement points. Results are in table 3.1.

As can be seen there are only small correlations between the personality factors within both groups at time 1. Only extraversion correlated with neuroticism and openness in both groups. Neuroticism also had a significant correlation with agreeableness in the EG. However, the size of the correlations was rather small. The picture remained the same within the CG at time 2. In fact, the correlations hardly differed. Within the EG, however, correlations between the personality factors increased dramatically. Only the correlations with openness remained about equal.

Thus, the results show that disregarding situational demand can lead to dramatically increased correlations between personality factors if the demand is high. It is very informative to inspect the variances in both groups. Within the CG the variance remained unchanged from time 1 to time 2. Within the EG, however, variance actually dropped. This is especially interesting in light of the fact that the correlations increased. Usually, a restriction in variance goes along with worse correlations. Here the opposite occurred. On the one hand, this is further evidence for the impact of situational demand on personality questionnaires. On the other hand, this might also be caused by a ceiling effect. Maybe the variance would have remained equal as well if the rating scale had been larger. Despite this, the increase in correlations in the face of diminishing variance is impressive.

Table 3.1 Correlations between personality factors

	N	E	O	A	C
time 1					
N	-	-0.42***	-0.11	0.09	-0.26
E	-0.32*	-	0.30*	-0.03	0.13
O	-0.13	0.36**	-	0.16	-0.02
A	-0.36**	0.15	0.18	-	0.03
C	-0.20	0.05	0.11	0.16	-
time 2					
N	-	-0.42***	-0.02	0.13	-0.28
E	-0.43***	-	0.29*	-0.06	0.11
O	-0.28	0.59***	-	0.24	-0.04
A	-0.39**	0.32*	0.27	-	0.03
C	-0.78***	0.46***	0.28	0.34**	-
SEM <sup>a</sup>					
N	-	.03	.04	-.41**	
E	-.05	-	.44**	.20	
A	-.42**	.85***	-	.03	
C	-.10	<-.01	.12	-	

Annotations. Below the diagonal are the correlations within the EG and above within the CG for time 1 and time 2, respectively. Significance levels have been Bonferroni corrected for 10 tests within each off-diagonal. <sup>a</sup> Significance levels have been omitted due to small sample size and violations of normal distribution.

The second step in the analysis was a multivariate analysis of variance with the independent variable group and repeated measurement factors for the personality factors. This analysis was followed by within and between group comparisons with a series of t-tests. All results as well as the descriptive statistics for both groups can be found in table 3.2.

The interaction between group and time was significant for all factors, openness being the exception. The between group analyses showed that the interaction was due to the fact that groups did not differ at time 1 but very much so at time 2. As was hypothesized, participants within the EG depicted themselves as less neurotic, more extraverted, agreeable, and conscientious. The effect sizes can be regarded as large. However, groups did not differ on the openness factor. Within each group, the change over time was as expected as well. Participants in the CG did not change from time 1 to time 2 except for a significant but slight drop in neuroticism and conscientiousness. Effect sizes were very small, though. Participants in the EG changed dramatically in their average personality factor scores. Openness again being the exception. As before, effect sizes can be regarded as large.

Table 3.2 Descriptive statistics and ANOVA results

	<i>M (SD)</i>		<i>interaction</i>	<i>between effect<sup>c</sup></i>		<i>within effect<sup>d</sup></i>	
	CG	EG		<i>t1</i>	<i>t2</i>	CG	EG
neuroticism t1	91.94 (25.99)	92.15 (24.85)	<i>F</i> 207.87***	<i>t1</i> -0.01 <sup>b</sup>	CG	0.14* <sup>b</sup>	
neuroticism t2	89.39 (24.56)	50.77 (20.49)	$\eta^2$ .53 <sup>a</sup>	<i>t2</i> 1.57*** <sup>a</sup>	EG	2.35*** <sup>a</sup>	
extraversion t1	116.12 (19.35)	121.45 (19.67)	<i>F</i> 29.12***	<i>t1</i> -0.27 <sup>b</sup>	CG	0.05 <sup>b</sup>	
extraversion t2	115.40 (18.16)	133.40 (14.03)	$\eta^2$ .14 <sup>a</sup>	<i>t2</i> -0.99*** <sup>a</sup>	EG	-0.86*** <sup>a</sup>	
openness t1	128.62 (15.35)	128.88 (17.99)	<i>F</i> 2.72	<i>t1</i> -0.02 <sup>b</sup>	CG	0.08 <sup>b</sup>	
openness t2	127.73 (16.57)	131.50 (14.09)	$\eta^2$ .02 <sup>b</sup>	<i>t2</i> -0.23 <sup>b</sup>	EG	-0.21 <sup>b</sup>	
agreeableness t1	117.50 (19.13)	116.33 (17.48)	<i>F</i> 31.62***	<i>t1</i> 0.06 <sup>b</sup>	CG	-0.10 <sup>b</sup>	
agreeableness t2	118.91 (19.51)	129.40 (13.21)	$\eta^2$ .15 <sup>a</sup>	<i>t2</i> -0.54*** <sup>a</sup>	EG	-1.06*** <sup>a</sup>	
conscientiousness t1	118.96 (20.10)	123.04 (19.31)	<i>F</i> 208.75***	<i>t1</i> -0.21 <sup>b</sup>	CG	0.13* <sup>b</sup>	
conscientiousness t2	116.97 (18.51)	153.82 (17.15)	$\eta^2$ .53 <sup>a</sup>	<i>t2</i> -1.99*** <sup>a</sup>	EG	-2.25*** <sup>a</sup>	
reasoning ( <i>g<sub>r</sub></i> )	114.15 (17.35)	118.77 (19.35)		<i>t1</i> -0.26 <sup>b</sup>			
LEWITE ( <i>g<sub>c</sub></i> )	.76 (.77)	.68 (.76)		<i>t1</i> 0.11 <sup>b</sup>			
LEWITE (SOE)	6.51 (2.32)	6.90 (2.22)		<i>t1</i> 0.17 <sup>b</sup>			
SEB	4.56 (4.94)	5.74 (4.47)		<i>t1</i> -.025 <sup>b</sup>			
statistics	44.54 (36.35)	53.26 (28.80)		<i>t1</i> -0.25 <sup>b</sup>			

Annotations.  $\eta^2$  = partial eta squared, significance levels for the interactions were Bonferroni corrected (5 tests), df (1,184); <sup>a</sup> power > .95; <sup>b</sup> power < .80; <sup>c</sup> displayed is Cohen's d (difference of the means divided by pooled SD in case of t1 and SD in CG in case of t2) as well as the Bonferroni corrected (10 tests) significance two-tailed for t1 and one-tailed for t2; <sup>d</sup> displayed is Cohen's d for repeated measurements (difference of the means divided by SD t1 and multiplied by  $\sqrt{2}$ ) as well as the Bonferroni corrected (10 tests) significance one-tailed for the EG and two-tailed for the CG; Maximum reasoning score possible was 210. PSE minimum score possible was -102, possible maximum was 102.

### 3.6.2. Controlling situational demand

Using structural equation models the following analyses controlled for the influence of situational demand on construct and criterion validity of the FFM. The multiple group model was designed as shown in figure 3.1. From the prior analyses it is obvious that openness was not faked. In order to reduce model complexity, openness was dropped from the following analyses. In order to achieve an acceptable model fit correlated errors between deliberation and angry hostility respectively straightforwardness had to be introduced to the model.

The model achieved an acceptable model fit:  $\chi^2$  [2089] = 3859.61, Bollen-Stine  $p$ -value = .32,  $SRMR$  = .139,  $RMSEA$  = .068 (90% CI .065 - .071),  $CFI$  = .803. According to the significance test the model cannot be rejected. However, the  $SRMR$  and the  $CFI$  indicate more connections not considered. Regarding the trait loadings an interesting result emerged in the EG. With the exception of the extraversion facet activity (E4), all other facets had significant trait loadings at time 1. However, at time 2 the trait loadings mostly dropped to insignificance in the EG. Conscientiousness had only one significant loading from achievement striving (C4), agreeableness from modesty (A5), neuroticism from anxiety (N1) and vulnerability (N6), and extraversion had four significant loadings, assertiveness and activity being the exceptions. A closer look at the loading pattern for extraversion revealed generally higher trait loadings at time 2 compared with time 1. Results for the control group are not reported because loadings on one factor were set equal. Regarding the state loadings within the CG, all loadings were significant except for straightforwardness (A2), modesty (A5) only at state 1, order (C2), and impulsivity (N5) which had no significant state loadings at both times. Within the EG the loading pattern was about equal at time 1, the only difference being additional

insignificant loadings for dutifulness (C3) and achievement striving (C4). In both groups loadings at time 1 were small to moderate and highest for extraversion facets. At time 2 the picture changed within the EG. With the exception of compliance (A4), modesty (A5), and excitement-seeking (E5B) all state loadings were significant. Loadings increased dramatically for neuroticism and conscientiousness facets (all  $a > .62$ ) and dropped for extraversion. The correlation between state 1 and state 2 amounted to  $r = .97$  ( $p < .001$ ) within the CG, but only to  $r = .27$  ( $p < .05$ ) within the EG. Obviously, the situational demand did not change within the CG, but it did change dramatically within the EG. Moreover, people in the EG responded differently to situational demand at time 1 and time 2, leaving only a small resemblance in rank.

Table 3.3 displays the means and variances for the latent variables derived from model 1. The table also contains the effect sizes for the latent mean comparisons. It can be seen that controlling for situational demand yielded only insignificant and meaningless group differences in all personality factor means. The only significant and moderate difference which occurred was for the mean of state 2 which was significantly higher in the EG. Both state variables had significant variances at both times and in both groups. The latter, however, cannot be said for extraversion which had no significant variance in the EG.

Of further interest was a comparison with a correlated trait model (model 2). Model 2 achieved the following fit:  $\chi^2$  [2078] = 3809.43, Bollen-Stine  $p$ -value = .34,  $SRMR = .131$ ,  $RMSEA = .067$  (90% CI .064 - .071),  $CFI = .807$ . Thus, the difference in CFI between both models was .04, indicating an improvement by letting personality dimensions correlate. The  $\chi^2$  - difference test yielded the following result  $\chi^2$  [11] = 50.18,  $p < .001$ . This also indicates a significant improvement from model 1 to model 2.

Table 3.2 Means and variances for latent variables

	M		$\sigma^2$	
	CG	EG	CG	EG
N	0	-0.02 (.00)	7.99***	21.63***
E	0	0.12 (.05)	3.97**	0.75
A	0	0.18 (.03)	8.27***	3.19*
C	0	0.48 (.08)	7.26***	4.16**
state 1	1.82	1.82	3.33**	4.74**
state 2	0	5.86*** (.78)	8.30***	6.71***

Annotations. \*\*\*  $p < .001$ , \*\*  $p < .01$ , \*  $p < .05$ . Means within the CG were fixed to be 0, thus, significant means in the EG represent significant group differences. An exception is the mean for state 1 which was set equal in both groups to express equal situational demand. Values in parenthesis are Cohen's  $d$ 's for latent means (see Hancock, 2001).

However, looking at the correlations (see table 3.1) reveals only small values, mostly close to zero, with only two significant correlations in each group. In the CG extraversion correlated with agreeableness and neuroticism with conscientiousness. In the EG it was also extraversion and agreeableness as well as agreeableness and neuroticism. A note of caution must be given at this point. Correlations between latent variables are corrected for attenuation and are therefore larger than bivariate correlations. Applying the equation from Hancock and Mueller (2001), yields construct reliabilities which allows to revoke corrections for attenuation. The construct reliabilities for neuroticism, conscientiousness, extraversion, and agreeableness in the CG were  $r_{tt} = .85$ ,  $r_{tt} = .89$ ,  $r_{tt} = .63$  and  $r_{tt} = .93$ , respectively. Within the EG extraversion reached a construct reliability of  $r_{tt} = .59$ , and neuroticism and agreeableness both reached  $r_{tt} = .83$ . Undoing the corrections for attenuation yields correlations between extraversion and agreeableness of  $r = .34$  within the CG and  $r = .59$  within the EG. The correlation between neuroticism and conscientiousness found in the CG decreases to  $r = -.36$ . And finally, the correlation

between agreeableness and neuroticism drops to  $r = -.35$  within the EG. All in all, the results indicate a slight advantage in model fit for the correlated trait model. However, the actual correlations found indicate almost unrelated personality dimensions.

The next step in the analyses was an investigation of the criterion validity of trait and state variable. The analysis was conducted with model 1 since the correlations between personality factors were only of secondary size. The only change to model 1 was that the manifest variable points in statistics examination (PSE) was introduced to the model and regressed on all personality factors as well as state 2. Table 3.4 contains the standardized loadings and the amount of variance explained for the latent regression (SEM) as well as the manifest regressions for both measurement times (t1 and t2).

Table 3.3 Regression coefficients

	CG			EG		
	t1	t2	SEM	t1	t2	SEM
N	.30*	.32*	.22	-.20	-.16	-.15
E	-.03	.04	-.04	-.16	-.13	-.15
A	-.08	-.03	-.02	-.22*	.04	-.26*
C	.08	.16	.16	.14	.03	.09
state 2	-	-	-.19	-	-	.09
$R^2$	.09	.09	.11	.09	.03 <sup>a</sup>	.13

Annotations. \*\*\*  $p < .001$ , \*\*  $p < .01$ , \*  $p < .05$ . t1 = Manifest regression with variables collected at time 1. t2 = Manifest regression with variables collected at time 2. SEM = Latent regression. <sup>a</sup> = correction for variance restriction (Buehner, 2006) results in  $R^2 = .09$ .

The results for the latent regression in table 3.4 indicate that the effect was comparable in both groups. However, the variance sources responsible for the effect

differed. In the CG it was mainly neuroticism which explained PSE. The direction indicated higher scores for those who described themselves as more afraid. Within the EG not only did this relationship switch here it was mainly agreeableness which explained criterion variance. The negative relationship indicates that less agreeable people achieved higher scores. This pattern is in line with the results from the manifest regressions analyses. The differences might be due to the unequal number of men in each group<sup>3</sup>. The other coefficients were comparable between both groups except for state 2. Within the CG state 2 had a negative loading which became positive in the EG. Moreover, the amount of explained variance strongly decreased in the EG at time 2 when situational demand was not controlled. This was caused by a restriction in variance. Correcting for this restriction rendered exactly the same amount of explained variance found at time 1. However, effects found in the manifest regressions were slightly smaller than in the latent regressions.

In a final analysis model 1 was used to explore the character of the state 2 variable in the EG. Therefore, the manifest variables reasoning ( $g_f$ ), lexical knowledge ( $g_c$ ), the self-overestimation score (SOE), and self efficacy beliefs of positive self-presentation (SEB) were introduced to the model and allowed to correlate with state 2. Results can be found in table 3.5.

Table 3.4 Correlations between state 2 and reasoning ( $g_f$ ), lexical knowledge ( $g_c$ ), self-overestimation (SOE), and self-efficacy beliefs (SEB)

	CG	EG
$g_f$	.07	-.08
$g_c$	.02	.13

<sup>3</sup> There were more men in the CG who were also less neurotic. Just looking at the men the correlation between neuroticism and PSE is positive and negative for the women. Since the group of men was very small no further investigations were conducted.

SOE	-.02	-.08
SEB	.07	.31**

Annotations. \*\*\*  $p < .001$ , \*\*  $p < .01$ , \*  $p < .05$ .

Within the CG state 2 had only very small and insignificant correlations with all four measures. However, in the EG there was a moderate and significant correlation with self efficacy of positive self – presentation beliefs. Of course, the same argument mentioned above, namely that the correlations might be overestimated due to correction for attenuation, holds true for these correlations as well. Yet, the correction was only for unreliability within state 2 and therefore of minor impact since state 2 had twenty four loadings, most of which were moderate to large and thus a construct reliability of  $r_{tt} = .96$  within the EG. This yields an uncorrected correlation of  $r = .30$  with SEB within the EG.

### 3.7. Discussion

Aim of the present study was to explore the impact of situational demand on the construct and criterion validity of a personality measure. It was hypothesized that situational demand impacts measures of personality structure and causes correlations between personality factors and that controlling for situational demand would render uncorrelated personality factors. Evidence for this hypothesis was provided. Furthermore, it was proposed that situational demand increases within a specific applicant setting compared to an average laboratory situation. This hypothesis could be confirmed. Regarding criterion validity of trait and state variance it could be shown that the state variable did contribute only little to the prediction when people faked. Thus, the conclusion that the actual criterion validity results from individual differences in personality traits can be drawn. Finally, the

results of the present study indicate that naming the additional variance state variance might be misleading. The character of the variable depends on the amount of situational demand: Low demand goes along with a state and high demand with a trait character. Moreover, the trait character can partly be explained by self efficacy of positive self – presentation beliefs.

In the introduction the most common critics on the FFM, namely, the number of factors, the methodological approach, orthogonality, and the impact of social desirability were listed. These critics mainly concern construct validity, but also criterion validity. Regarding the number of factors, the present study offers no new insights since this was not an aim. However, much can be added to the arguments dealing with the latter three problems.

### *3.7.1. Construct validity*

The present analysis used a LST design and structural equation modeling. Structural equation models are insightful since they offer actual tests for models. Unlike in exploratory factor analysis, the model test helps to find out whether a proposed model fits the data or not. Another methodological advantage of the present study is the LST design. It allows to control for situational demand. The result is quite astonishing. Not controlling for situational demand leaves large differences in personality factors between CG and EG. However, all these differences can be led back to the situational demand, yielding groups which no longer differ. Thus, a LST design might be helpful whenever a situational impact is likely. This means virtually always, since we never assess people in a situational vacuum (Deinzer et al.).

Regarding the problem of orthogonality, it was hypothesized that correlations between personality factors are merely due to situational demand or state. The present results provide evidence for this hypothesis, but also show that even when situational demand was controlled, a few correlations remained. Thus, a lack in discriminant validity as suggested by Costa and McCrae (1992b) cannot totally be ruled out. The *SRMR* underlines this problem since it indicates further connections between the facets. However, considering that most correlations between personality factors were insignificant when situational demand was controlled, the lack might not concern all factors. Most of all extraversion and agreeableness seem to be problematic since they still shared variance in both groups even when situational demand was controlled. Extraversion, furthermore, had another peculiarity. Its pattern loading on the state and trait variables was more dependent on situational demand than was the case for any of the other personality factors. In the EG the state loadings were smaller when people faked than under honest conditions. In other words, when situational demand was low extraversion loaded higher on the state variable indicating that it contains more state variance than other factors. However, when people faked and the character of the state variable changed, the loadings also decreased. This highlights findings by Rost, Carstensen, and von Davier (1999) who conducted analyses for unidimensionality according to item response theory within the data from the German normative sample for the NEO-FFI and found extraversion to be a substantially heterogeneous construct.

However, usually it is the correlation between neuroticism and conscientiousness which raises concerns and constantly is the highest correlation reported (Mount, et al., 2005; Pauls & Crost, 2005b; Toomela, 2003). This correlation was also the highest in the present study when situational demand was high. Controlling for situational demand not only decreased the correlation, it fell to

virtual unimportance. All in all, there is a lot of evidence for the hypothesis by Ziegler, Bühner, and Toomela that correlations between personality factors are not only caused by a lack in discriminant validity, but more so by situational demand. Orthogonality might, after all, be a question of situational demand.

The last concern described above deals with the impact of social desirability. Two contradicting lines of research were presented above. Smith and Ellingson (2002) concluded that situational demand (social desirability) does not impact construct validity. On the other side were Ziegler, Bühner, and Toomela suggesting the opposite. Both used structural equation models to underline their claims. However, neither of them applied a LST model. This model was realized within the present study, allowing a clear separation of trait and situational demand variance. The results show that situational demand impacts construct validity of measures of personality structure. Within a faking condition, correlations between manifest variables increased. This increase could be reversed controlling for situational demand. Moreover, means for all personality factors, except openness, were significantly larger under a faking instruction compared to an honest condition as well as a control group. Controlling for situational demand, the CG and the EG no longer differed in any of the personality factors. Thus, it could be shown that situational demand (social desirability) or state impacts relations between personality factors.

### *3.7.2. Substance vs. style*

In the first part of this paper it was mentioned that there is a debate about the character of faking: is it an ability or is it social desirability? In order to shed light onto this question, the variable state 2 was correlated with two ability and two trait

measures. In the CG there were no correlations with any of these measures. State 2 was mainly marked by neuroticism facets. The negative regression weight which occurred for the prediction of points in a statistic examination shows that it might be aspects like test anxiety which drive this variable. Within the EG the character of state 2 was different. There was a significant correlation with self efficacy of positive self – presentation beliefs. In other words, the belief to be able to impress others was correlated with state 2. Moreover, state 2 had a positive regression weight in the prediction of points in a statistics examination. Thus, people who believe in themselves achieved higher scores. Furthermore, SEB was negatively and moderately correlated with neuroticism. Thus, one can conclude that state 2 also reflects individual differences in neuroticism. However, while it is test anxiety in the CG, it is the opposite, namely, having no fear and high self confidence in the EG. It has to be noted, though, that SEB accounted only for a small portion of faking variance. Thus, there is room for other variables to explain faking as well.

All in all, the present analysis lends support to the findings by Ones, Viswesvaran, and Reiss that faking reflects actual individual differences in personality traits. However, there is a switch in the character of the situational demand variable (state 2): Depending on the amount of situational demand it is more like test anxiety or more like a self efficacy variable. Thus, the term social desirability does not capture the whole spectrum of possible meanings of the variable and might in fact be misleading. Thus, the question should not be substance versus style, but which substance?

### *3.7.3. Criterion validity*

In the introduction the question was raised whether the predictive power of faked personality questionnaires results from trait or faking variance. The present results show that the contribution of faking variance is only very small. Thus, it must be real individual differences in personality traits which cause criterion validity. However, the results also show that only very few facets have significant loadings on the personality traits under faking conditions. This means that the criterion validity does not result from the personality trait as a whole, but from specific facets. The present analysis revealed that for the chosen criterion it were low modesty and low depression for example. This is in line with recent findings which favor the use of the facets instead of the factors when predicting performance (Lounsbury, Sundstrom, Loveland & Gibson, 2002). Ones, Visweswaran, and Reiss (1996) came to the conclusion that social desirability does not affect criterion validity. The present results indicate that the contribution of faking variance is indeed small. However, criterion validity is affected since faking decreases personality to a few facets.

### *3.7.4. A hypothesis regarding faking*

From the present results it could be hypothesized that a mediation took place. People do not fake everything as the present analysis confirmed. Thus, faking requires a decision on whether to fake or not. This decision can be based on specific knowledge: It could be assumed that people base their judgment of what to fake on specific knowledge regarding the circumstances of the situation. This means, in our case, if I know what qualities are expected of a psychology student, I

can portray myself in that way. However, faking is connected to the confidence of being able to live up to the distorted picture drawn as can be seen by the correlation between state 2 and SEB. Thus, the hypothesis drawn from this study is that SEB mediates the relationship between specific knowledge and amount of faking. Unfortunately, lexical knowledge as assessed in the present study is no good indicator for such specific knowledge and the hypothesis cannot be tested. This mediation would be in line with part of the model of faking suggested by McFarland and Ryan (2000). They hypothesized that ability to fake, beliefs toward faking, opportunity to fake, and situational influences (e.g., desire for job, warning of Lie scale) could be important factors predicting faking behaviour on personality tests. At least as far as beliefs towards faking are concerned, the present study supports that model. Nevertheless, further research is still needed here.

#### *3.7.5. Limitations*

The biggest limitation of the present study is the relatively small sample size given the complexity of the models. Sample size influences the accuracy of parameter estimations. Thus, a replication with larger samples is needed.

A replication study will also give information on the correlated errors which were added here. If they cannot be replicated, they were nothing but sample peculiarities in the present study.

Furthermore, it could be argued that a study using students and a statistics examination bears only little practical relevance. Of course, testing the hypotheses in applicant settings would be most interesting. However, as Rogers (1997) stated, the sample should have practical relevance. Since the whole experimental design was customized for a university setting, this practical relevance is given. However,

this practical relevance is limited to a university setting. Within a workplace setting like an applicant setting different results regarding factors faked and facets affected by faking might occur. The only drawback of the present sample is the fact that it is partly pre – selected for intelligence. All psychology students have been selected for their grades in school. A certain cutoff had to be passed. Thus, a more homogeneous group regarding intelligence occurred. This might have affected the correlation between reasoning and state 2.

Summing up, the present study clearly showed an impact of situational demand on construct as well as criterion validity of a personality structure questionnaire. Controlling for situational demand not only deleted effects of faking on personality means, it also yielded an almost uncorrelated personality model. Moreover, controlling for situational demand recovered regression coefficients which occurred under honest conditions. Unfortunately, the question whether faking is a personality trait or an ability could not be answered. The present results point into the direction of personality traits.

Thus, the current item set used by Costa and McCrae certainly is a “serviceable model” (Costa and McCrae, 1992a), but it might even be closer to the final orthogonal factor solution of personality than believed.

**Chapter 4: Situational Demand, Response Sets, and Faking in the  
Light of Cognitive Interviews and Mixed Rasch Models**

## **Chapter 4. Situational Demand, Response Sets, and Faking in the Light of Cognitive Interviews and Mixed Rasch Models**

Psychological tests are one of the most commonly used methods in selection processes. A lot of research is concerned with the predictive validity of these tests (e.g., Schmidt & Hunter, 1998). Psychological tests can roughly be divided into subjective self assessments and objective ability tests. A prominent representative of the former are personality questionnaires. The research on the predictive validity of personality questionnaires is sizeable (e.g., Barrick, et al., 2001; Chamorro Premuzic & Furnham, 2003; Furnham & Chamorro-Premuzic, 2004; Furnham, et al., 2002; Salgado, 2002; Salgado, 2003). All in all, results suggest that using personality questionnaires is a good way to improve the predictive validity of selection processes. However, as long as such questionnaires exist there have been doubts regarding their susceptibility towards faking. Three research lines concerned with faking can be distinguished. There are those researchers who try to find out whether faking impacts criterion validity of questionnaires (e.g., Hough, Eaton, Dunnette, Kamp & et al., 1990; Schmitt & Oswald, 2006; e.g., Viswesvaran & Ones, 1999). The meta-analysis by Viswesvaran and Ones came to the conclusion that faking impacts mean scores but not criterion validity. Ziegler and Bühner recently demonstrated this effect (submitted). Using sophisticated methodological models they could show that the variance due to faking contributes only little to the prediction of a real life criterion. Even though those results indicate that the overall criterion validity is not affected by faking there are also research results which demonstrated that faking does affect the ranking of people.

Hough (1998) could show that participants who faked more ranked higher. This underlines the hypothesis proposed by Ziegler and Bühner that faking reduces the number of predictive facets to just a selected few. This in combination with the little added information gained from faking might cause changes in ranking.

A second line of interest is the impact of faking on the construct validity of questionnaires (Ellingson, et al., 2001; Pauls & Crost, 2004; Pauls & Crost, 2005b; Schmit & Ryan, 1993; Smith, et al., 2001; Topping & O'Gorman, 1997; Ziegler & Bühner, submitted; Ziegler, et al., submitted-b). While results in the first line of research all pointed in one direction, the picture here is less clear. Some researchers, like Smith and Ellingson, concluded that neither social desirability nor applicant settings impact the construct validity of personality questionnaires. In the same work mentioned above, Ziegler and Bühner could show that this conclusion was premature. They showed that situational demand influences the correlations between personality factors and thus, construct validity.

The third line of research is concerned with the question: What is faking? So far one thing is clear, people differ in the way they fake and do not just endorse the highest categories (e.g., Rosse, et al., 1998; Ziegler & Bühner, submitted; Ziegler, et al., submitted-b). Rosse and colleagues could show that there are extreme fakers, slight fakers, and even honest participants within an applicant sample. This finding was confirmed by other researchers using an item response theory approach (Zickar & Robie, 1999a; Zickar, et al., 2004; Zickar & Robie, 1999b). Thus, there is evidence that faking is performed differently by people. McFarland and Ryan (2000) proposed a model of faking, linking it to several variables such as ability to fake, beliefs about

faking and situational circumstances. However, the empirical base of this model is scarce. Pauls and Crost (Pauls & Crost, 2005a) could show that faking can be predicted by cognitive ability and self efficacy of positive self – presentation beliefs. The latter has been confirmed by Ziegler and Bühner. However, a shortcoming of those studies was that it was believed that all participants would act the same when situational demand is high. Consequently, all fakers were treated as one group. The results by Zickar, Gibby, and Robie imply that this is not appropriate.

Summing up, there is evidence that situational demand causes faking which does not impact criterion validity, but construct validity of personality questionnaires. Moreover, responding to situational demand is not a homogeneous response set. Consequently, the present paper aimed at four goals. The first aim was an investigation of the psychological processes which occur when questionnaires are faked. Secondly, the replication of the number of response classes reported before was intended. Thirdly, it was tried to find out why people differ in the faking strategy they use. And finally, the criterion validity found in the different faking classes were compared. Within the following paragraphs the four goals are described in more detail.

## 4.1. The present study

### 4.1.1. *Psychological processes*

In order to identify the psychological processes happening when people intentionally distort their answers to a personality questionnaire, a qualitative analysis technique, the cognitive interview, was used (Dillman & Redline, 2004; Fowler, 1995; Willis, 2004). Two different cognitive interview types can be distinguished: the concurrent and the retrospective method. The concurrent method, also called think aloud technique, requires participants to express any thought crossing their mind when working on the questionnaire. These thoughts are recorded in writing as well as on tape. Later, the thought protocols can be analyzed to find strategies. The second method, the retrospective cognitive interview, takes place after participants filled out their questionnaire. Using a structured interview, participants are then asked questions regarding cognitive processes which occurred during questionnaire completion. Usually, cognitive interviews are conducted to improve legibility of questionnaires. However, the method can also be used to identify different strategies for distorting a questionnaire. In the present study, both techniques were applied on a sample of participants asked to fake a personality questionnaire. The retrospective interview asked participants directly for their strategy. The answer could later be tested using the information from the concurrent cognitive interview.

#### *4.1.2. Replication of response classes*

As mentioned above, using an item response theory approach three different faking classes have been identified: regular responders (rr), slight fakers (sf), and extreme fakers (ef). The method used to find these classes is called Mixed Rasch Models (MRM) and was developed by Rost (e.g., Rost, 1985; e.g., Rost, 1991; Rost, et al., 1997). The MRM basically is the fusion of Latent Class Models and Rasch Models. Thus, it allows classes which differ qualitatively in the measured trait or ability. However, unlike in a Latent Class Model people in one class can differ quantitatively in the trait or ability. In other words, MRMs look for the number of qualitatively different classes in each of which the Rasch Model holds true. Moreover, these classes are supposed to describe best the answer patterns observed. Rost, Carstensen, and von Davier (1999) also were the first to use MRMs in personality research. Analysing the data from the German normative sample of the NEO – FFI, they found two classes of people, those who endorsed middle rating categories with a higher probability and those who preferred extreme rating categories. This result has been replicated by Austin, Deary, and Egan (2006) in a British sample. Thus, even when situational demand is low, there are two different classes of people. This fact can be attributed to different response sets. However, if situational demand increases, the number of classes goes up to three. Now, two different faking response sets can be distinguished. There are slight fakers and extreme fakers. Finally, an honest response group can be identified. What are possible reasons for these classes? The present study will use data from the qualitative analysis as well as data collected by Ziegler and Bühner in a

different study to find answers to this question. One of the advantages of that study was its known group design (Rogers, 1997) which allows to put people in one of two groups: faker (F) and non-faker (NF). Participants from the qualitative analysis were asked to fake and belong in group F. The number of different classes within the combined sample will be investigated. Afterwards, the distribution of the classes in the two groups can be analyzed.

#### *4.1.3. Causes for differences in responding to situational demand*

McFarland and Ryan (2000) proposed a model of faking. So far, cognitive ability and self efficacy of positive self – presentation beliefs (SEB) have proven to be connected to faking (Pauls & Crost, 2005a; Ziegler & Bühner, submitted). However, different faking strategies were not differentiated. Within the present study this will be done using MRMs. In order to find causes for the different faking styles, class membership will be predicted by cognitive ability (fluid and crystallized), SEB, self overestimation (SOE), age, gender, and personality traits assessed under no specific situational demand and thus fairly honest. Ziegler and Bühner conducted the personality questionnaire twice starting with an honest condition for each participant.

#### *4.1.4. Criterion validity*

As far as we know, no research on differences in the criterion validity within different faking classes exists. This issue will be explored within the study as well.

After the faking classes will have been found, separate regressions of a statistics examination score on the personality scores assessed will be conducted and compared. Moreover, a new way of correcting for faking will be explored. Using MRMs not only allows to find different classes, it also corrects the person parameter according to class membership. Thus, the person parameter of an extreme faker will be corrected for faking. The corrected person parameters will also be used in separate regressions and results will be compared with the results derived from the uncorrected values under honest and faking conditions. Ferrando and Chico (2001) showed that ordinary lie scales can detect faking better than person fit indices within a two parameter logistic model (2PLM). The present analyses will also investigate the ability of MRMs to detect faking. Since a known group (Rogers, 1997) design was used, the accuracy of MRMs in detecting faking can be assessed. Compared with 2PLMs, MRMs offer the advantages of the Rasch model, especially specific objectivity and are thus, preferred within the present analyses. Moreover, Ziegler and Bühner found that individual differences in responding to situational demand contributed positively to the prediction of points in a statistic examination. Using class membership according to the MRM results as predictor and controlling for intelligence, it will be tested whether class membership and thus, individual differences in faking techniques contribute to the prediction of points in the statistics examination.

Furthermore, the ranking of the corrected personality parameters will be compared with the ranking found under honest and faking conditions. If the correction of the personality parameters is good, the ranking should be highly comparable to the honest condition and less so to the faking condition.

## 4.2. Method

As mentioned above, the current study consists of two parts: a qualitative and a quantitative analysis. The qualitative analysis is the first part and was conducted in a sample of  $N = 50$ , especially drawn for that purpose. Aim of the qualitative analysis was to find out which different strategies are used to intentionally distort a personality questionnaire. The second part of the study, dealing with the MRM analyses and the exploration of the resulting classes, will use the data from the qualitative analysis as well as from a previous study (Ziegler & Bühner, submitted). Thus, the combined sample had a size of  $N = 236$ . Detailed descriptions regarding sample and procedure for the latter data can be found there. Thus, the following sample and procedure description is only concerned with the sample from the qualitative analysis.

### 4.2.1. *Sample and procedure*

The sample consisted of 50 (34 women and 16 men) students enrolled in different faculties at the Ludwig-Maximilians-University in Munich. However, 27 of them were psychology students. The average age and semester were 22.26 ( $SD = 1.91$ ) and 1.89 ( $SD = 2.01$ ), respectively.

All participants had to fill out the personality questionnaire twice. When working on it the first time, participants were given the same fake good instruction used in the previous study. Thus, they were told to imagine themselves in a selection scenario: The following test would be used to select psychology students and only good

performance would assure selection. The second time was without a special instruction. Participants were asked to answer as honestly as possible. While working on the questionnaire for the first time, participants were further instructed to express any thoughts aloud. An approval for taping was asked and given by all participants. An experienced test expert also took notes. Afterwards participants were administered a semi – structured interview. They were asked how they tried to achieve the given goal, whether they applied that technique to all questions, and finally if they could name the strategy they had used. After the interview participants completed the personality questionnaire for a second time. Due to the think aloud technique participants were administered the tests one at a time using a computer. Since this was very time consuming, it was not possible to administer the cognitive ability tests or any of the other measures used by Ziegler and Bühner besides the personality questionnaire. Unfortunately, the sample consisted of students from higher semesters, which made it impossible to use their statistic examination grades and thus, the sample in the quantitative analyses regarding criterion validity only consisted of participants from the previous study ( $N = 186$ ).

All fifty interviews and thought protocols were analysed by two people whose task it was to find the different strategies and assign each participant to one of them. To assess the degree of convergence between the raters, Cohen's Kappa was computed.

#### 4.2.2. Test materials

The NEO – PI – R (Costa & McCrae, 1992b) was used as personality measure. It consists of 240 items. Eight items always measure one of the six facets for each of the five factors of the FFM, i.e., neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness. Items ask participants to rate themselves in typical behaviors or reactions on a five-point Likert scale ranging from “strongly disagree” to “strongly agree”. Alphas for the factors range from .87 to .92. For the facets they range from .53 to .85.

The basic module from the Intelligence Structure Test (IST) (Amthauer, et al., 2001) was used to measure fluid intelligence (reasoning). Three out of nine subtests combine to measure verbal intelligence, numerical intelligence, and figural intelligence, respectively. The combination yields a reasoning score (Cronbach  $\alpha = .97$ ). The theoretical basis for the reasoning measure (gf) has been shown (Beauducel, et al., 2001).

The Lexical Knowledge Test (Lexikon-Wissen Test, LEWITE, Wagner-Menghin, 1998) was applied as a measure of crystallized intelligence. The LEWITE is an adaptive test measuring crystallized intelligence ( $g_c$ ) (Cronbach  $\alpha = .86$ ). Participants complete sentences which define difficult words. They also have to state whether they know the word or not, which allows the computation of a self – over – estimation score (SOE) reflecting how much participants overestimate their own ability.

Furthermore, the Efficacy of positive Self-Presentation Questionnaire (Mielke, 1990) was applied. It assesses how much participants believe to be able to make others

think they are smart, likeable and capable. Thirty three items have to be rated on a four point Likert scale ranging from “totally disagree” to “totally agree”. The questionnaire has been described in more detail by Pauls and Crost. As those authors suggested a total score for self efficacy beliefs of positive self – presentation (SEB) was computed. Cronbach  $\alpha$  (on item level) of this measure was .86.

As criterion variable served the score from a statistics examination<sup>4</sup>. The grade has been used in other studies as well (Furnham & Chamorro-Premuzic, 2004) and proved to be predictable by personality measures. The statistic examination consisted of 33 multiple choice items with an internal consistency of  $\alpha = .80$ . All participants took the examination at least 2 months after being subjected to the other tests used in the study.

#### 4.2.3. Statistical analysis

Analyses were conducted using SPSS 14.0 and WINMIRA 2001 for the MRMs. As far as the qualitative analysis is concerned, only one statistic, Cohen’s Kappa, was computed to assess rater agreement.

As mentioned above, MRMs were conducted to identify the number of classes which best describe the data. In the current analysis MRMs were only conducted for the six conscientiousness facets. There are two reasons for this. First of all, conscientiousness has proven to be a good predictor of performance across different jobs and criteria (Barrick, et al., 2001; Chamorro Premuzic, et al., 2004; LaHuis, et al.,

---

<sup>4</sup> 24 % of the participants did not take part in the statistic examination. Their score was estimated using SPSS 14.0 and the EM method. In order to get better estimates, scores from the tests taken between the personality questionnaires were also used for estimation

2005; Robertson, Baron, Gibbons, MacIver & Nyfield, 2000; Salgado, 2003). Secondly, since the number of classes is only one of the topics, it would be beyond the scope of the present paper to investigate results for all five personality factors. Usually, item answers are the basis for analysis. However, within the present study conscientiousness as a whole was to be investigated. The NEO-PI-R combines six different facets as a measure of conscientiousness. The items for these six facets surely measure one higher order factor, but also a lower order factor. It would be impossible to fit a MRM to all items since they do not measure exactly the same trait or combination of traits. Thus, a different approach was chosen. Item answers per facet were averaged and the six mean facet scores were used in the analyses. These mean facet scores reflect the answer category chosen on average for each facet. Models ranging from one class to four classes were tested. Model fit was judged by two model test statistics: Cressie Read (C.R.) and Pearson  $\chi^2$  ( $P^2$ ). The test statistics can be interpreted like  $p$  – values in a significance test. The null hypothesis is that the model fits the data. Thus, a value below .05 indicates that the chosen number of classes does not fit the data. Both test statistics only assess model fit accurately if each possible answer pattern is observed at least once. The combined sample consisted of  $N = 236$  participants. Given the five possible categories for each of the six mean facet scores, the number of patterns to observe would be 3072. However, only 93 different patterns were observed. Thus, the test statistics were evaluated using a bootstrap procedure with 200 samples (Buehner, 2006). Other authors have used information criteria to find the best model (Austin, et al., 2006; Zickar, et al., 2004). While information criteria find the best model of a given number of models they do not

assure that this model fits the data at all. Thus, they were not used for decisions in the present study.

In order to predict class membership by the variables mentioned above, a multinomial logistic regression was conducted. The logistic regression was preferred over a discriminant analysis because it makes no assumptions regarding the distribution of the independent variables (Howell, 2002, p. 583). Criterion validities were assessed with a linear regression. Unfortunately, not all participants took the statistics examination and were also part of the faking group. Thus, the sample for the linear regressions consisted of only  $N = 92$ .

Since reporting and discussing results for the two different parts would not improve the understanding of the results, both parts will be separated. At first results from the qualitative analysis will be reported and discussed. Afterwards, results from the quantitative analyses will be reported followed by a general discussion.

### 4.3. Qualitative Analysis

#### 4.3.1. Results

After analyzing the cognitive interviews and the thought protocols the two judges independently of each other concluded that there were two main strategies for intentionally distorting the questionnaire. The strategies were named *slight faking* and *extreme faking* in accordance to Zickar et al (2004). Cohen's Kappa for rater agreement was .77 which can be regarded as good. About 20 percent of the participants used the second strategy and the majority of 80 percent only faked slightly. When analysing the data the first step in the analysis was the answer of the participant on the direct question for a strategy. In most cases this proved to be of little merit since participants could not really name a strategy. Of more help were the participants' answers to the question how they had tried to achieve the goal of faking good. Here the two strategies emerged. Most of the people said something like they took their real answer and pushed it a little in the right direction. These answers were confirmed by the actual thought protocols. To give an example, one participant said the following when pondering his answer to question 40 (I keep my things clean and tidy.): "Well, I guess I don't really do that, but if I want to be selected I better endorse a four." Only few participants, those labelled as extreme fakers, stated that they endorsed the highest possible category. However, one important aspect which needs to be mentioned is that participants did not fake all items regardless of their content. Before considering an answer, participants judged whether such a question would

reveal information important to select psychology students. If that was the case, they faked. If it was not the case, participants did one of two things. They either answered honestly, or they answered neutrally, some even did both alternatingly.

Another interesting finding could be observed. Students enrolled in psychology faked other aspects than students from other faculties. The differences were rather severe. While psychology students tended to endorse conscientiousness items and portray themselves as low in neuroticism, other faculty students faked openness and portrayed themselves as more neurotic. These observations were tested using a Multivariate Analysis of Variance with faculty as independent and the faked Big 5 scores as dependent variables. The result showed a significant and large effect ( $F[5;44] = 3.07; p = .02; \eta^2 = .26; Power = .82$ ). Post-hoc tests with Bonferroni corrections revealed significant differences in neuroticism and conscientiousness. The effect sizes were  $d = 1.00$  ( $p < .01; Power = .97$ ) and  $d = -.61$  ( $p < .05; Power = .69$ ), respectively. These effect sizes are the difference in group means divided by the standard deviation of the psychology students. Positive values indicate a higher mean for students from other faculties. Thus, psychology students faked significantly and moderately to largely more in neuroticism and conscientiousness.

#### 4.3.2. Discussion

The results from the qualitative analyses confirmed the quantitative evidence for individual differences in faking. Some people have no fear of being detected and simply endorse the maximum category while other people are more careful. This is in

line with findings by Zickar and colleagues (Zickar, et al., 2004; Zickar & Robie, 1999b) who had reported those faking styles in their quantitative analysis of a combined applicant and incumbent sample. Even the class sizes reported there are in line with the class sizes found in the present analysis. However, analyzing data from a faking experiment with an honest and two different fake good groups, Zickar et al. (2004) only reported two different classes. This might have methodological reasons since a different questionnaire was used than in the other analysis. The authors themselves do not offer any more specific explanations why they found two classes in one analysis and three in the other.

Moreover, the results also lend support to the hypothesis that people do not simply fake all items when situational demand is high (e.g., Pauls & Crost, 2005b; Ziegler & Bühner, submitted; Ziegler, et al., submitted-b). People first evaluate the importance of an item in terms of the situational demand (e.g., application for a certain job or student program). If the item has no importance regarding the demand of the situation, they do not fake. As a consequence, even the use of the maximum response strategy does not necessarily result in the maximum score. This result bears two questions. First, what enables people to judge the importance of an item? Ziegler and Bühner (submitted) have proposed that it is specific knowledge, e.g., about the job applied for. The current results underline this hypothesis, since participants always considered what would be good for a psychology student.

Another result showed that students with a psychological background faked differently from students with no psychological background. Other faculty students portrayed themselves as more neurotic and less conscientious. This was not only found

by the two raters, but was also confirmed by a significance test. This fact not only underlines the importance of specific knowledge it also shows that implicit theories might be used by those lacking that knowledge. Thus, judgement of item importance can be based on specific knowledge or implicit theories about the requirements for a psychology student.

All in all, the qualitative analyses confirmed individual differences in the way people intentionally distort a questionnaire. The psychological process of faking is not as easy as some might have believed. The actual answer is derived after an evaluation of item importance which is followed by either an aggravation of one's own score or the endorsement of the maximum category, respectively. The evaluation of item importance must at least in part be based on specific knowledge.

#### 4.4. Quantitative Analysis

The next paragraphs report the results from the quantitative analyses. The order is in accordance with the issues raised in the introduction.

##### 4.4.1. Replication of faking classes

The results of the MRM analyses for the six mean facet scores of conscientiousness are presented in table 4.1.

Table 4.1 Results of MRM analyses.

number of classes	Cressie Read	Pearson $\chi^2$	CAIC
1	<.001	<.001	2145.50
2	.01	<.001	2300.88
<b>3</b>	<b>.14</b>	<b>.08</b>	<b>2509.84</b>
4	.02	.02	2763.31

Annotations: Values below .05 indicate model misfit. The best solution is printed bold

From table 4.1 it can be seen that the solution with three classes fitted the data while none of the other solutions did. Thus, the sample consisting of fakers and non fakers can best be described with three classes. The first class contained 59 percent of the participants, the second 25.2 percent and the third 15.8 percent. The estimated mean probabilities of assigning a person to the correct class were .92 for class 1, .93

for class 2, and .90 for the third class. The probabilities of assigning a person to a wrong class were all below .10. Thus, class assignment can be regarded as accurate. Table 4.2 contains the percentage of fakers and non fakers in each of the classes as well as the descriptive statistics for the other variables. The anova reliabilities estimated by WINMIRA were .81 (class 1), .78 (class 2), and .90 (class 3) and can be regarded as satisfactory.

It can be seen that most fakers belong to class 1. However, most non fakers do so as well. Nevertheless, a big part of the non fakers belongs to class 2, while only very few fakers were assigned to that class. Regarding class 3 it was the other way around. Looking at the faked conscientiousness score, class 1 can be regarded as the slight faking class, class 2 as the regular response class, and class 3 as the extreme faking class. Table 4.2 also contains comparisons between the three groups. It can be seen that all groups differed significantly in the various conscientiousness measures with values increasing with amount of faking. Effect sizes can be regarded as large with one exception. The honest conscientiousness scores of the two faking groups differed only slightly.

Interestingly, both faking groups differed in the other honest personality factor scores, reasoning, and their self efficacy of positive self – presentation beliefs (SEB). Effect sizes were moderate and the direction indicated higher values for the extreme faking group. The differences between the extreme faking group and the regular respondents also reached moderate effect sizes in these variables, the direction being the same. However, both groups were rather small and thus, power was not sufficient to yield significant results.

Table 4.2. Descriptive statistics for the three classes.

class	1 (sf)	2 (rr)	3 (ef)	sf vs. rr	sf vs. ef	rr vs. ef
non faker	48 (51.1%)	42 (44.7%)	3 (4.3%)	-	-	-
faker	81 (69.2%)	5 (4.3%)	32 (26.5%)	-	-	-
theta	1.66 (1.94)	-1.22 (1.64)	5.72 (3.94)	0.91***	-1.29***	-2.20***
faked C <sup>a</sup>	148.67 (12.89)	114.20 (19.02)	168.61 (9.49)	2.01***	-1.16***	-3.18***
honest C	124.41 (18.41)	108.13 (17.40)	127.49 (22.85)	0.81***	-0.15**	-0.96***
honest N	93.25 (25.91)	91.40 (24.82)	79.46 (21.59)	0.07	0.54*	0.47
honest E	116.68 (20.14)	118.60 (17.39)	127.46 (18.86)	-0.10	-0.55**	-0.45
honest O	124.47 (17.79)	129.51 (15.62)	133.20 (20.76)	-0.28	-0.48*	-0.20
honest A	116.54 (18.91)	114.40 (17.16)	123.66 (13.28)	0.12	-0.40	-0.52
g <sub>r</sub>	114.82 (18.62)	115.78 (19.11)	122.47 (15.86)	-0.05	-0.42	-0.36
g <sub>c</sub>	0.68 (.75)	0.79 (.75)	0.76 (.82)	-0.14	-0.10	0.05
SOE	6.80 (2.24)	6.76 (2.57)	6.31 (1.89)	0.02	0.21	0.20
SEB	4.54 (4.99)	5.07 (4.45)	7.29 (3.52)	-0.11	-0.58*	-0.47
age	22.81 (5.44)	24.80 (7.32)	21.28 (3.39)	-0.35	0.26	0.61*
women in %	77.5	74.5	62.9	-	-	-

Annotations: Displayed are means and standard deviations (in brackets) for all continuous variables and percentages for categorical variables. sf = slight faking; rr = regular response; ef = extreme faking. The last three columns contain Cohen's *d* (difference in group means divided by standard deviation derived from the total sample) and Bonferroni corrected significance levels (\*\*\*) =  $p < .001$ ; \*\* =  $p < .01$ ; \* =  $p < .05$ ) for the continuous variables. The significance tests were only conducted with participants from the first study by Ziegler and Bühner, group sizes were:  $N = 129$  (class 1),  $N = 45$  (class 2), and  $N = 32$  (class 3). Power was above .80 for effect sizes larger than .50 (sf vs. ef), .43 (sf vs. rr), and .59 (ef vs. rr). g<sub>r</sub> = reasoning, g<sub>c</sub> = crystallized intelligence, SOE = self – overestimation, SEB = self efficacy of positive self – presentation beliefs. <sup>a</sup> All statistics were derived from the total sample minus the  $N = 94$  participants from the first study who were part of the control group.

The only further observed difference in the continuous variables occurred in age which decreased with faking. Kruskal – Wallis – tests were conducted for the categorical variables faker / non faker and gender. The differences in gender between the classes did not reach significance ( $\chi^2[2] = 3.09$ , n.s.) while the difference in faker / non faker reached significance ( $\chi^2[2] = 56.30$ ,  $p < .001$ ). Thus, from the differences in the descriptive statistics it can be concluded that more reasoning, extraversion, more openness, less neuroticism, and more self efficacy of positive self – presentation beliefs lead to an extreme faking style and not to a slight faking style. The slight faking style mostly rests on a higher base in the honest trait (C) compared with the regular responders.

Using a repeated measure ANOVA with class as independent variable and conscientiousness (honest and faked) as repeated factor, the within group changes from honest to faked answers were tested. The analysis yielded a significant and moderate interaction between class and measurement time ( $F [2;114] = 5.75$ ;  $p = .004$ ;  $\eta^2 = .092$ ;  $Power = .90$ ). Post – hoc tests revealed effect sizes of  $d = .88$  for slight faker and  $d = 1.45$  for extreme faker between both measurement times. The effect size represents the difference between time 2 (faked) and time 1 (honest) divided by the honest standard deviation. Within the regular response group no difference occurred ( $d = -.01$ ).

#### 4.4.2. Causes for differences in responding to situational demand

In order to further investigate causes for the different response sets, a multinomial logistic regression with class membership as dependent variable and class 2 (rr) as reference category was conducted. In a first step all predictors were used and the results were hard to interpret due to the large number of predictor variables. Most predictors had very small regression coefficients. In order to improve interpretability, the predictors with the smallest regression coefficients ( $<.01$ ) were dropped from the analysis. The resulting model fitted the data according to the Likelihood – ratio – test ( $\chi^2 [16] = 59.28, p < .001$ ). Nagelkerke's  $R^2$  was .32, indicating a moderate effect. Table 4.3 contains the overall significance for each predictor as well as regression coefficients. The regression coefficients reflect changes in the log odds of being a member of one of the faking classes compared to the regular respondent class. Negative values indicate that the log odds of being a member of one of the faking classes go down if the value of the predictor goes up.

Table 4.3. Results of the multinomial logistic regression with class 2 (rr) as reference category.

	$\chi^2$	$B (1)$	$B (2)$	$Wald \chi^2 (1)$	$Wald \chi^2 (2)$
A	10.15*	-.005	.037*	.17	4.97
C	26.07***	.054***	.047*	21.12	10.10
g <sub>f</sub>	1.67	-.004	.014*	.12	.68
g <sub>c</sub>	.63	-.228	-.240	.58	.37
SOE	1.81	-.030	-.152	.11	1.55
SEB	12.04*	-.081	.091	2.69	1.98
age	5.89	-.066	-.133	3.79	3.29
gender	2.25	.242	.876	.25	2.09

Annotations: B (1) contains the regression coefficients indicating changes in log odds of being in class 1 and not 2, the same holds true for B (2). Degrees of freedom for all  $\chi^2$  tests were 2 and for the Wald  $\chi^2$  tests 1. A = Agreeableness (honest), C = Conscientiousness (honest) for names for the other variables see table 1. \*\*\* =  $p < .001$ ; \*\* =  $p < .01$ ; \* =  $p < .05$

Table 4.3 reveals that agreeableness, conscientiousness, and self efficacy of positive self – presentation beliefs served as significant predictors for class assignment. Regarding the classification as an extreme faker higher agreeableness, conscientiousness, reasoning, and SEB increased the log odds. Furthermore being young and being male increased the log odds of classification as extreme faker. For the slight faking class it was mostly higher conscientiousness which contributed to class assignment. All other variables decreased the log odds of being assigned to the slight faking group. However, sample size and the rather large number of predictors decreased power. Thus, it might be worth to consider all coefficients regardless of their significance.

The next step in the analysis was a comparison of criterion validities for the honest, the faked, and the MRM conscientiousness scores. Points in a statistics examination served as criterion. Moreover, in a second step class membership was tested as a predictor.

#### *4.4.3. Criterion validity*

During the analysis of the classes it could be seen that they differed in regard to their mean scores of conscientiousness in the honest, the faked, and also in the MRM scores. The differences were large except for the difference in honest conscientiousness between the two faking groups. In other words, fakers on average did not differ much in their conscientiousness, but the differences in faking styles yielded large differences in faked group means. Thus, criterion validities for the three

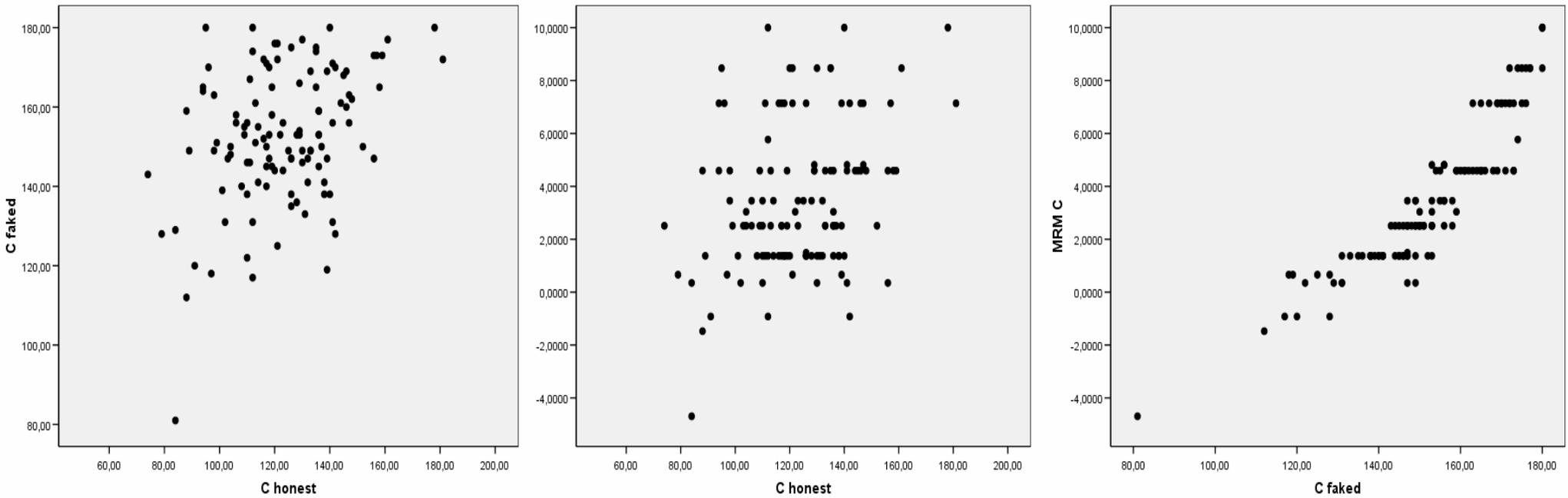
scores were computed. The resulting  $R^2$ 's were .01 for the honest conscientiousness score, .02 for the faked conscientiousness score, and .03 for the MRM conscientiousness score. Neither of the values reached significance, which can be attributed to a lack of power which was below .40 in all three cases. The smaller value for the honest conscientiousness score can be attributed to a smaller variance within the score. Thus, the criterion validities for the three different scores do not differ.

In order to test the predictive validity of individual differences in response style, the points in the statistics examination were also regressed on class membership. In a second step reasoning was added to control for intelligence. Class membership explained 2.4 percent ( $p < .05$ ) of the variance. Reasoning added 14.2 percent ( $p < .001$ ). The standardized regression coefficient for class membership decreased from .17 ( $p < .05$ ) to .12 ( $p < .09$ ) when reasoning was added. Comparing the means in the statistics examination for the different classes revealed the highest scores for the extreme faking group. The difference to the slight faking class was  $d = .50$  ( $p < .05$ ;  $Power = .77$ ), and to the regular response class  $d = .48$  (n.s. ;  $Power = .27$ ). The slight faking and the regular response class did not differ  $d = .08$  (n.s. ;  $Power = .29$ ). Thus, class membership contributed to individual differences in points in the statistics examination: extreme fakers achieved higher scores. Moreover, this contribution was only to a smaller part due to reasoning.

Another impact of faking can be seen in the ranking of participants. In order to test whether MRM scores reconstruct the ranking seen under honest conditions, the three different rankings were compared. To give a quick overview three scatter plots for the different conscientiousness score combinations were drawn (See figure 4.1).

As can easily be seen from the plots, the ranking in the honest condition differs strongly from both other conscientiousness scores. The faked score and the MRM score, however, result in largely comparable rankings. The results from the visual inspections were confirmed by a look at the three Spearman's rhos which were .30\*\* (honest and faked), .27\*\* (honest and MRM), and .94\*\* (faked and MRM). A closer look at the ten most conscientious people according to the honest condition revealed that only two of them remained within the top ten in the faking scores and MRM scores. Regarding the top twenty, six of them remained in the top twenty in the faking scores and four in the MRM scores. Thus, the ranking found under honest conditions differs strongly from that found under faking condition and cannot be reconstructed through the corrections performed within the MRM analyses.

Figure 4.1 Scatter plots of the rankings



Legend. C honest = Honest conscientiousness cores; C faked = faked conscientiousness scores; MRM C = MRM corrected conscientiousness scores

#### 4.5. General Discussion

The present analyses had four goals which all aimed at understanding more of the psychological process taking place when people intentionally distort their answers to a questionnaire. In other words, the study was supposed to investigate what happens when people react to situational demand. Moreover, possible causes for individual differences in responding to situational demand were investigated. Finally, it was explored whether there were any differences in the criterion validities of different conscientiousness scores. The results of the qualitative and the quantitative analyses show that the psychological process happening when people fake a questionnaire starts with the evaluation of the importance of an item regarding the situational demand. This evaluation is based on specific knowledge, implicit theories, and reasoning. The actual answer then depends on the preference for one of three response sets: regular response (rr), slight faking (sf), and extreme faking (ef). People with high self efficacy of positive self – presentation beliefs tend to endorse extreme categories. Moreover, the actual answer has a solid base in the real conscientiousness score. Finally, it was shown that honest, faked, and MRM corrected conscientiousness scores did not differ in their criterion validity, but in the ranking of people. In the following paragraphs the different response classes will be discussed and a model of responding to situational demand will be proposed. Moreover, the practical implications of the other findings will be discussed.

#### 4.5.1. *Response classes*

In line with Zickar et al. (2004) three different response classes were found: An extreme faking class, a slight faking class, and a regular response class. The results of the comparison between the three classes show that people with higher neuroticism only faked slightly. This is probably due to the fact that these participants were more afraid of being caught with a lie. They did not want to exaggerate their answers too much, probably fearing not being able to live up to that standard. The role of neuroticism in faking has already been mentioned elsewhere (Viswesvaran, et al., 2001). However, more important for the classification as a slight faker is the reported honest trait. In other words, people who have higher conscientiousness scores when responding honest also have a higher probability of being assigned to the slight faking class when intentionally distorting the answers. Thus, within the slight faking group the score has a solid base in the real trait. Within the qualitative analyses it could be seen that slight fakers add one or two points to the answer they would give under honest conditions. Extreme fakers on the other side are characterized by high self efficacy of positive self – presentation beliefs. They also have the highest honest conscientiousness scores. Thus, the extreme response set might be due to a high starting point in some cases – there just was not enough rating scale left. Yet, it is also reasonable to assume that the extreme faking was caused by the belief to be able to impress others (SEB) and by the ability to find out which items must be faked to impress most. This latter ability is partly due to reasoning ability. However, the qualitative analysis also showed that participants use their knowledge and implicit theories to evaluate the importance of an item. Pauls and Crost (Pauls & Crost, 2005a) reported similar results, but did not differentiate between different faking styles. They argued

that the self efficacy beliefs gave people the courage to fake and reasoning the ability. Under the light of the present results, this line of argumentation has to be restricted to the extreme faking group. Within the slight faking group there were participants with smaller reasoning ability and self efficacy beliefs. These participants also fared worse in the statistic examination compared with the extreme faking class. Thus, it is reasonable to assume that slight fakers did not only lack the courage to fake extremely, but also the ability. However, even the slight faking group achieved larger faked conscientiousness scores compared with the regular response group and also with their own honest scores. Summing up this evidence, it can be concluded that extreme faking is triggered by higher self efficacy of positive self – presentation beliefs, higher honest trait scores, and cognitive abilities. Slight faking on the other hand results from honest trait scores which are gently exaggerated into the right direction. This slight exaggeration is in part due to higher neuroticism but also to smaller reasoning. Ziegler and Bühner hypothesized that faking comprises specific knowledge about the goal of situational demand. They based that hypothesis on the finding that a situational demand variable contributed a little, but positively to the prediction of a real life criterion. Within that analysis they did not differentiate between the different response sets reported in the present study. Thus, slight fakers and extreme fakers were regarded as one. From the present analyses it is clear that reasoning increases the log odds for classification as extreme faker and decreases them for classification as slight faker. In other words, reasoning had an inverted relationship to extreme faking compared with slight faking. This might explain why Ziegler and Bühner found no correlation with reasoning. Thus, the original hypothesis by Ziegler and Bühner must be expanded and reasoning must be added to the list of personality characteristics contributing to individual differences in responding to situational

demand. A reasonable possibility is that people use knowledge to judge an item and hence fake it. However, if they do not possess this knowledge, they make inferences using their reasoning ability. Disturbing is the result that a higher crystallized intelligence increased the log odds of being assigned to the regular response class and not one of the faking classes. At first glance, this contradicts the hypothesis. However, the qualitative analysis showed that the evaluation of an item must be based on knowledge since participants used their knowledge of the necessities for a psychology degree. Moreover, knowledge was operationalized as lexical knowledge within the present study. Lexical knowledge certainly does not contain a lot of information regarding psychology studies. Moreover, the actual differences in lexical knowledge between the different response sets were small and thus, the results in the logistic regression might be attributable to chance. The following paragraph sums up these results and proposes a model for responding to situational demand.

#### *4.5.2. Model of responding to situational demand*

As mentioned above the actual answer to an item in a personality questionnaire depends on the situation (situational demand high or low) and on attributes and traits of the person responding (e.g., age, gender, SEB,  $g_f$ ,  $g_c$ ) which result in a preference for a certain response set. This psychological process of evaluation based on knowledge and followed by an interaction of traits and attributes of a person is evidenced by the present results. Figure 4.2 illustrates a model of responding to situational demand.

Figure 4.2 Model of Responding to Situational Demand

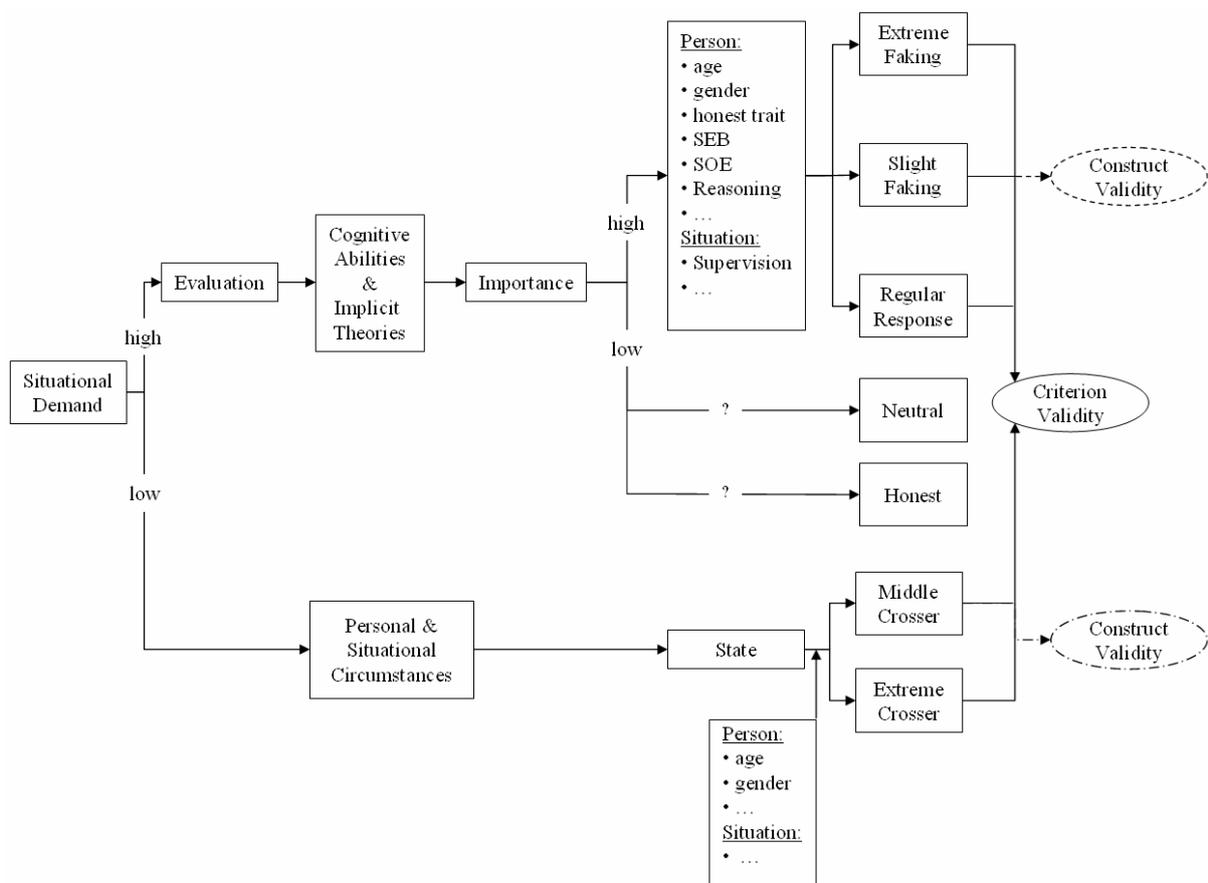


Figure 4.2 consists of two different major paths. The lower path proposes the psychological processes occurring when situational demand is low. As reported by Ziegler and Bühner this results in a state of the person which influences responses to the questionnaire. Possible influences can be traits like test anxiety and social desirability. This state has its base in the personal and situational circumstances. Within a laboratory setting state anxiety is probably higher than in an everyday situation. Moreover, the result in a personality questionnaire would certainly be different when a person was out partying last night or has just been left by his or her partner. Such state variance can account for up to twenty percent of the variance in a personality questionnaire (Deinzer et al., 1995). The only difference which still seems to occur even under normal conditions is the preference for a certain response set (Austin, et al., 2006; Rost, et al., 1999). Some people prefer

middle categories and others extreme categories. Austin and colleagues reported that men and younger people prefer more extreme categories. However, a qualitative analysis as was conducted here would shed more light onto this part of the process. The upper path of the figure shows what happens if situational demand is high. People judge the items based on their specific knowledge, implicit theories, and reasoning and choose a response set based on certain characteristics of the person and the situation. Thus, cognitive ability, personality, and situational characteristics determine the preference for a response set. Under high situational demand only five percent of the participants who were asked to fake still used a regular response style. It may be that these persons did not follow the instructions or were just too honest to fake. The vast majority, however, either used a slight or an extreme faking response set. Possible causes for the difference might lie in the cognitive ability and in the real personality traits of a person as discussed above. Another possibility is that the response set preferred under normal conditions caused the faking response set. Middle crosser might prefer a slight faking and extreme crosser an extreme faking response set. The connections between the different response sets under the different situational demands should be investigated further. As McFarland and Ryan proposed in their model of faking, the opportunity to fake also plays a role within the process. This is acknowledged by the situational influence on the preference for a response set. Opportunity to fake might also include aspects such as the announcement of cross checking (Supervision). However, one of the main differences between slight and extreme fakers are the higher beliefs to be able to impress others within the latter class.

Another result of the present study and the prior study by Ziegler and Bühner is the impact of situational demand on the validity of personality measures. Both studies demonstrated that criterion validity is not affected by situational demand as

far as the overall effect size is concerned. This is indicated in the model through one criterion validity to which both paths lead. However, construct validity is directly affected because situational demand causes correlations between personality factors. This is highlighted in the model by two different construct validity variables. The construct validities not only differ in terms of correlations between personality factors, but also in the composition of variance which indirectly affects criterion validity. This is discussed in the next paragraph.

#### *4.5.3. Practical implications*

Ziegler and Bühner could show that the response to situational demand served as a positive, but small predictor for a real life criterion when situational demand was high. They hypothesized that the predictive validity stems from knowledge and self efficacy beliefs. The present results lend support to this hypothesis and add reasoning to the list of contributors. Furthermore, it also shows that the preference for a response set affects the overall score and thus, the ranking. Moreover, extreme fakers performed better in the criterion making class membership a valid predictor as well. Thus, even though criterion validity remains the same under honest and faking conditions, the sources of predictive power differ. This is underlined by three aspects. First of all, the changes in ranking which occur under faking conditions indicate different compositions. Secondly, the faking variable within the study by Ziegler and Bühner served as a positive predictor. And thirdly, class membership also served as a valid predictor with higher scores for extreme fakers. Thus, if one selects personnel using personality questionnaires, the base for selection still consists partly of true differences in personality, but also of differences in cognitive ability and other traits like self efficacy of positive self – presentation beliefs. Some

people might want to circumvent this by correcting for faking. The present study indicates that using MRM corrected scores does not help very much. Previous findings showed that lie scales detect faking better than IRT models (Ferrando & Chico, 2001). Ferrando and Chico used a two parameter logistic model (2PLM) and person fit indices to detect fakers. The present analyses show that using Mixed Rasch Models (MRM) helps to detect fakers with an exceptional accuracy of about 95 percent. However, using MRMs to correct for faking proved less useful. This might be due to the small number of facets used. With more items the estimation of the person parameters is more accurate and more differentiated. Thus, under these circumstances MRM corrections might prove more helpful. Other methods for the correction of faking have been proposed as well. For example lie scales (e.g., Eysenck & Eysenck, 1964), ipsative measures (e.g., Baron, 1996; Bowen, et al., 2002; Martin, et al., 2002), objective measures (McClelland, et al., 1989), cut off scores (Ziegler, et al., submitted-a), and many others as well. Ziegler et al. discussed the different methods and came to the conclusion that using objective measures probably is the best way to guard against faking. Yet, the question remains whether one needs to guard against faking. The present results indicate that faking might not be such a problem after all. Of course, the purity of the measure is lost, however, personality traits remain the main contributors to the criterion validity. And, moreover, new and most importantly valid information is gained. After all, the individual differences affecting faking seem to help to achieve goals in real life. Having higher self confidence and cognitive abilities certainly helps to achieve goals in life. Thus, future research should investigate the type of knowledge responsible in evaluating item importance and how it contributes to success in life.

#### 4.5.4. Methodological implications

As described above, other researchers have used information criteria to find the best fitting MRM model (e.g., Zickar, et al., 2004). These criteria take the number of parameters to estimate into account and also use the likelihood of the data. Thus, the model with the smallest number of parameters and an acceptable likelihood is chosen as the best. However, these criteria do not test whether the specific assumptions of the Rasch model are actually fulfilled. Thus, even though the information criterion might be the smallest, the model might not fit. This is underlined by the present results. According to the CAIC a 1 class model would have fitted the data best. However, the model tests show that only the 3 class model really fits the data. Thus, it is suggested to use a combination of both. Model tests should be used to find the models which actually fit and information criteria can be looked at to find the best model of those fitting.

#### 4.6. Limitations

The present analyses used complex methodological models which require large samples. The sample used here certainly was not small, but it might not have been large enough either. Thus, a replication of the results is necessary. Moreover, the results are limited to the special setting.

Furthermore, it could be shown that the number of items used in a MRM affect chances of finding all possible response sets (Zickar & Burnfield, 2003). Within the present study six items representing the six facets of conscientiousness have been used. While this is a small number, the results found are in line with the results by Zickar et al. (2004) who also reported three response classes.

Of course, one can criticise the fact that the sample used consisted of students only. This might be a threat to the external validity of the experiment. However, in accordance to the suggestions by Rogers (1997) the faking scenario was realistic for participants, ensuring external validity. Nevertheless, a replication of the current results within a real life setting such as an applicant setting would broaden the base of the proposed model.

Summing up, the present analyses demonstrated that situational demand causes a psychological process involving cognitive abilities as well as personal and situational characteristics resulting in a specific response set. Thus, scores in faked personality questionnaires not only contain actual personality traits, but also other sources of variance. However, these sources of variance contribute to the prediction of performance. So the answer to the question of whether to use personality questionnaires for selection should be: "Yes, but beware of the content!"

## **Chapter 5 : Conclusions**

## Chapter 5. Conclusions

### 5.1. Summaries of the three studies presented

#### 5.1.1. Study 1

Study 1 reanalyzed data collected by Toomela (2003). The data consisted of personality self ratings and cognitive ability test results from  $N = 912$  men with military background. In his original article Toomela showed that in the group with the highest cognitive ability neuroticism and conscientiousness had substantial cross loadings and could no longer be clearly separated using exploratory factor analysis. The present reanalysis was based on the hypothesis that situational demand was responsible for this. It was assumed that situational demand was a systematic error which adds itself to all personality facets affected by faking and causes correlations between these factors to increase. In a first analysis step it was explored why the different ability groups responded differently to the situation. In other words, why did the highest cognitive ability group respond more intensely to the situational demand than any of the other groups? During this analysis it was hypothesized that the situational demand was felt due to military rank and not due to intelligence. This was assumed since the men in the highest military group mostly came from the group of officers. The group of non commissioned officers mostly portrayed average ability and finally, the group of recruits below average ability. Drawing from these results it was reasoned that officers responded more to the situational demand because they felt obliged to draw a prototypical military picture of themselves. While this might have occurred within the non commissioned officer group as well, it certainly was not felt by recruits who only do their required

military service time. All in all, it was assumed that situational demand was felt highest by the officer group, followed by the non commissioned officers, and finally the recruits. Using a multigroup structural equation model with aspects of Latent State Trait theory the hypothesis could be confirmed. Moreover, the results indicate that an uncorrelated trait model might represent personality better when situational demand was controlled. The officer group had the highest means in the latent variable containing the situational demand variance. Both other groups followed as was hypothesized. The control of situational demand strongly decreased the correlations between the personality factors. Thus, first evidence that situational demand interferes with the orthogonality of personality factors was provided. However, since the data only contained measures from one measurement point, the hypothesis could not be confirmed terminally. Therefore, study 2 was conducted using an experimental design.

### 5.1.2. Study 2

In study 2  $N = 186$  participants were randomly assigned to one of two groups. The first group was a control group which twice filled out a personality questionnaire honestly. Participants of the other group, the experimental group, were given specific faking instructions at the second time. Between both measurements the groups worked on a reasoning test, a test of crystallized intelligence, and also a questionnaire of self efficacy of positive self presentation beliefs.

This design was chosen for two reasons. First of all, the two measurement points allow the separation of questionnaire variance into trait and state variance. Thus, the influences of state variance on the correlations between personality factors could be investigated. As in study 1 it was hypothesized that controlling for situational demand would strongly diminish the correlations between personality factors. The second reason concerned criterion validity. Due to the separation of state and trait variance it was also possible to distinguish trait and faking variance. Thus, it could be explored whether the predictive power of personality questionnaires in applied settings is a result of trait or faking variance. Moreover, the ability to fake, represented by the individual differences in the latent state variable fake, were correlated with the other assessed measures in order to find out what enables people to fake. The two differing views that faking is an ability versus that faking is a personality trait were compared.

The results again gave evidence to the hypothesis that situational demand is responsible for correlations between personality factors. Just looking at the bivariate correlations between the personality factors showed substantial correlations in both groups. However, when people faked these correlations strongly increased as was

hypothesized. When situational demand was controlled using structural equation modeling the correlations in both groups strongly decreased. Most correlations were close to zero when situational demand was controlled. Yet, extraversion and agreeableness still correlated substantially. Thus, it was concluded that situational demand is one of the causes for the correlations between personality factors. However, the lack in discriminant validity, claimed by Costa and McCrae (1995), cannot totally be ruled out even though it can be limited to two specific personality factors.

As a further result the faking variable revealed the strongest relationship with self efficacy of positive self – presentation beliefs. None of the other variables shared variance with faking. Thus, it was concluded that faking is promoted by the belief to be able to impress others.

Finally, it could be shown that individual differences in faking added positively, but very little to the prediction of a real life criterion. The personality traits, however, had the lion's share of the explained variance. Drawing from these results, it was hypothesized that faking must also rely on specific knowledge which helps to be successful. More precisely, it was assumed that knowledge is necessary to fake, however, only those who also believe to be able to impress others actually do fake.

One of the short comings of study 2 was that faking was seen as a homogeneous response set. No individual differences in faking strategies were assumed. In order to find out whether such strategies exist and to shed more light onto the psychological process of intentionally faking in a questionnaire, study 3 was undertaken.

### 5.1.3. Study 3

Study 3 actually consisted of two different parts. In the first part of study 3 a sample of  $N = 50$  participants was asked to fake a personality questionnaire using the same specific instruction as in study 2. Moreover, participants were asked to express any thoughts aloud. This was followed by an interview about faking strategies and another filling out of the questionnaire under honest conditions. The analysis of the thought protocols and the interviews revealed two different faking strategies. Common to both strategies was that not all items were faked. Only those answers to those items which were perceived as important regarding the faking goal were distorted. The faking style, though, differed. In one group people only increased their honest score slightly. This group, which made up eighty percent of the sample was called slight fakers. The other twenty percent chose the most extreme answer category to portray themselves. Unimportant items were either answered honestly or by using the middle category. Another interesting finding was that participants who were not enrolled at the psychology faculty faked different personality aspects. They portrayed themselves as more neurotic and less conscientious than students from the psychology faculty. Thus, the qualitative analysis of study 3 provided evidence for the assumption that specific knowledge is part of faking. Moreover, it also showed that implicit theories also play an important role.

The second part of study 3 used MRMs to analyze the combined data from study 2 and 3. MRMs were used to find different response classes. The personality factor conscientiousness was chosen for the analysis. In accordance with previous research findings three different response sets could be identified: extreme faking, slight faking, and regular response. Participants which were part of the control

group in study 2 make up the bigger part of the latter class. Fakers from study 2 as well as study 3 were mostly assigned to one of the faking classes. Thus, the findings of the qualitative analysis were confirmed.

In a series of further analyses it could be shown that the slight fakers were also more neurotic than extreme fakers. Moreover, it could be shown that honest conscientiousness scores built a solid starting point in both faking classes. Extreme fakers also showed the highest values in self efficacy of positive self – presentation beliefs. An interesting finding was that extreme fakers also had a larger reasoning ability. Thus, it was concluded that slight fakers are not only more afraid of being caught, they also lack the ability to fake and do not believe they could live up to an overly exaggerated picture of themselves. Extreme fakers on the other side have high self efficacy beliefs and reasoning ability. Thus, reasoning was added to the list of contributors to faking.

Moreover, it was shown that membership in one of the classes served as a valid predictor for a real life criterion. Members of the extreme faking class fared better. Controlling for the higher reasoning ability, class membership still contributed to the prediction.

Regarding the comparison of criterion validities in the different classes, no differences could be observed. Thus, criterion validity can be regarded as unaffected by different response sets. However, the ranking of participants under honest and faked conditions strongly differed. MRM corrections were also not able to reconstruct honest scores. This means that even though criterion validity remains unaffected by situational demand, the ranking is strongly affected. This bears consequences for the construct validity.

Drawing on the different results, a model of responding to situational demand was proposed. The model allows for two different response paths depending on the

situational demand. Low situational demand either leads to middle crossing or to extreme crossing. Mediating variables might be situational circumstances as well as age and gender. High situational demand on the other hand leads to one of the three response sets reported above. Personal as well as situational variables are proposed as mediators. Finally, it was assumed that both paths lead to the same criterion validity, but have different construct validities.

## 5.2. Conclusions regarding the goals of the present project

In the introduction four different goals of the present project were deduced. These goals were pursued with three studies whose results were described and discussed above. In the following four sections each goal will be revisited and the various results will be integrated. Finally, a short summary and an outlook for future research are provided..

### *5.2.1. Impact of Situational Demand on the Construct Validity of the BIG 5*

The Big 5 have been proposed as a theoretical model consisting of five orthogonal factors (Costa & McCrae, 1992b; Goldberg, 1992). However, there is a lot of empirical evidence that the factors have substantial correlations (Mount, et al., 2005). Costa and McCrae attribute this fact to a lack in discriminant validity for some of the facets. Other possible explanations have been presented in the introduction.

Within the present project it was hypothesized that situational demand in the form of systematic error variance causes those correlations. In this sense situational demand would influence the response given by a person. This influence would impact all items which are susceptible to the specific situational demand. Accordingly, correlations between those items would increase. Evidence for this hypothesis was provided.

Using Latent State Trait theory this situational demand was separated from the trait variance. Correlations derived under these conditions were smaller than correlations without controlling for situational demand. However, a few correlations between the personality factors remained. Specifically, extraversion and

agreeableness still shared variance. However, the other correlations strongly decreased when situational demand was controlled.

This contradicts conclusions drawn by Ellingson et al. (2001) as well as Smith and Ellingson (2002). Both researcher teams had reported that situational demand has only little influence on the construct validity of personality questionnaires. However, they only used one measurement point and were therefore not able to clearly distinguish between state and trait variance. Moreover, they allowed personality factors to correlate. When they compared applicant and student samples and concluded that they did not differ, they only consulted the loading patterns, though. A look at the intercorrelations reveals different patterns with larger correlations in the applicant setting. Thus, this shortcoming in combination with the results from the present studies let a different conclusion appear more likely. Situational demand influences construct validity of personality questionnaires: Correlations between personality factors inflate with increasing situational demand.

Another interesting finding was that higher situational demand changes the ranking of people in a questionnaire compared with honest results. This has already been reported by Hough (1998). However, he did not consider the implications for construct validity. Differences in ranking denote differences in the composition of the variance under high demand conditions from that under honest conditions. One could argue now that this change is due to individual differences in faking and thus, in cognitive ability or another personality trait. However, the present results also allow for a different explanation. When situational demand increases and people fake, personality factors lose variance. This can be explained by the increasing state variance. A consequence of this variance loss is that a personality factor just has loadings from a few selected facets. Here people still differ and thus, the facets have variance which allows for higher loadings. Consequently, changes in ranking might

be attributable to these facets. In other words, a broad personality factor no longer represents six facets as under honest conditions, but only one or two. Within these facets the ranking might differ strongly from the ranking in the factor. This, in combination with the individual differences in faking and faking style, leads to a changed ranking.

All in all, the present project revealed an impact of situational demand on the construct validity of a personality questionnaire. Not only does it cause correlations between personality factors, it also changes the variance composition of a given personality factor.

### *5.2.2. Impact of Situational Demand on the Criterion Validity of the BIG 5*

If situational demand impacts the construct validity, it is likely to assume that it also affects the criterion validity of a personality questionnaire. In line with previous findings (Ones, et al., 1996; Viswesvaran, et al., 2001) the present studies show that this is not the case. The variance of a real life criterion explained by a personality questionnaire remains about equal no matter whether honest, faked, or corrected scores are used.

However, it was also shown that individual differences in faking add little, but positively to the prediction of a real life criterion. Moreover, the present analyses explored such possible individual differences in faking and showed that different response sets are used to fake. People with higher cognitive ability, less fear, and more self efficacy beliefs fake extremely. Less bright and more fearful people fake only slightly. Thus, the differences in faking style represent differences in important personality aspects. These aspects are certainly important in real life and might explain why faked questionnaires keep their predictive validity. However, this

conclusion would be premature. Despite the evidence for the predictive power of individual differences in faking, the present results also demonstrate that personality traits contribute most strongly to the amount of explained variance. Thus, it is not faking, but personality plus a little faking which explains variance. In the previous section it was explicated that faked personality questionnaires have different construct validity. The factor variance no longer comes evenly from the facets, but is highly determined by a few facets. Taking this into consideration, the conclusion for the criterion validity is that it does in deed not result from the personality factors, but from the facets which determine it. Thus, criterion validity would be due to only a few facets. Recent studies have shown that using facets instead of personality factors might help to explain more variance (Lounsbury, et al., 2002). Future research should try to explore which facets keep their variance and are thus, the most likely contributors to criterion validity. Furthermore, it should be investigated whether the same facets are the only contributors to criterion validity under honest conditions. While this is highly likely, a contradiction of that result would be highly interesting since it would mean that different personality aspects contribute to criterion validity depending on the assessment situation. An explanation for this would be that life also consists of different situations and thus, different personality aspects might influence the behavior of a person leading to a more or less successful goal achievement.

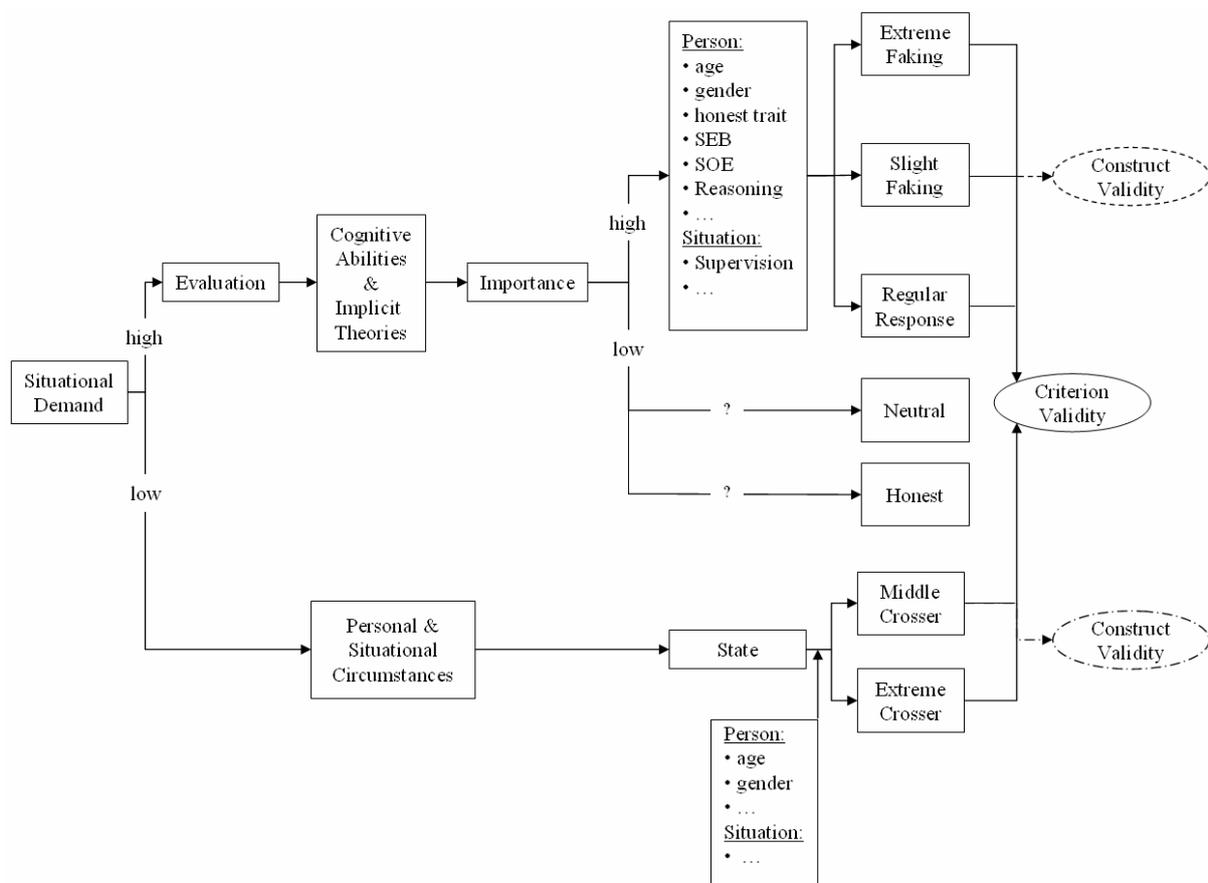
All in all, the present results indicate that situational demand has no direct impact on the criterion validity of a personality questionnaire. However, an indirect impact could be demonstrated. Criterion validity no longer results from a broad factor, but only from a few selected facets. Faking ability adds only little.

### 5.2.3. *The Psychological Process behind Situational Demand*

Especially the qualitative analysis helped to shed more light onto the psychological process of intentionally distorting answers to a personality questionnaire. These results in combination with the other results led to the proposition of a model of responding to situational demand.

The model (figure 5.1) distinguishes between high and low situational demand. Low demand represents ordinary assessment situations. Such situations are characterized by absolute anonymity and bear no consequences for the participant. Despite this, there is evidence that situational influences account for up to twenty percent of the variance (Deinzer, et al., 1995). Thus, situational circumstances influence the answering behavior. Moreover, Rost, Carstensen, and von Davier (1997) could show that under normal conditions two response sets can be found: Middle and extreme crosser. In a recent paper Austin et al. (2006) showed that men and younger people preferred extreme categories. Thus, besides situational influences, personal characteristics also influence the answer. The result is one of the two response sets. Since most of the research regarding the development of the Big 5 was conducted with students in laboratories under low situational demand conditions (John, et al., 1988), construct validity should equal the results from most of these studies. Criterion validity has also been investigated under these circumstances (Furnham & Chamorro-Premuzic, 2004; Furnham, et al., 2002) and was found to be comparable with results derived in applicant settings with higher situational demand.

Figure 5.1 Model of responding to situational demand



The psychological process occurring under higher situational demand is depicted in the upper path of the proposed model. From the present results it can be concluded that people first judge the importance of an item. Cognitive abilities like reasoning and specific knowledge are used as well as implicit theories for this judgment. Importance in this case refers to the achievement of a certain goal which caused the situational demand. Such goals can range from simple impression management to getting a job. If an item is assumed to be unimportant, people either answer honestly or neutrally. The mediating variables remain unclear so far. If the item is important in the eye of the participant, one of three different response sets is chosen. Very few people choose to answer honestly, despite the situational demand. The majority fake slightly and some participants fake extremely. Neuroticism, self efficacy beliefs, reasoning, and the faked trait itself mediate this as well as situational variables like the announcement of a test expert checking the results.

Furthermore, the results show that even under these conditions the criterion validity remains the same. However, construct validity is different as was explained above. The proposed model also contains two question marks. During the qualitative research it became clear that participants either answered honestly or neutrally to items they believed to be unimportant. However, it was not possible to determine reasons for these different answer styles. It is reasonable to assume that the trust the participants had into the anonymity of the research project determined the answer style.

All in all, the proposed model based on the results of the three studies depicts the psychological process and its consequences which occur when people respond to situational demand.

#### *5.2.4. Dealing with Situational Demand*

One of the goals of the present study was to test a new approach for the correction of faking. In study 3 MRMs were used to determine the number of different faking classes. However, one of the advantages of MRMs is that they also estimate person parameters for all participants which are located on the same trait variable. This means that the person parameters from different classes can be compared. Since MRMs consider the class attributes in the parameter estimations, the parameters are corrected for the response set. In other words, extreme fakers get parameters which are corrected towards the middle and slight fakers get parameters corrected away from the middle. Thus, it might be possible that MRMs reconstruct the trait as it can be observed under honest condition.

The results of the present analysis showed that MRMs are extremely sensitive to faking. That means fakers were assigned to one of the faking classes with a high

security. While this is an encouraging result, the corrected scores performed as bad as the uncorrected faking scores. Moreover, the corrected scores were closer to the faked scores than to the honest scores. This result does not advocate the use of MRMs for the correction of faking. However, this conclusion is not justified. Within the present analysis only six items were used. Thus, the parameter estimates were not as accurate as they would be with more items. Moreover, such a small number of items limits the range of possible parameters. All members of one class who have the same sum score are assigned the same person parameter. Thus, the variation of parameter scores is closely linked to the number of items. Therefore, it should be tested whether a larger number of items is more able to reconstruct the honest score.

Before closing this section, it has to be pointed out that the results of the present studies imply another conclusion. It was shown that the criterion validity of the personality questionnaire did not change when faked. Moreover, the biggest share of explained variance belonged to the trait and only a small part to the faking variance. Furthermore, the faking variance proved to be a positive predictor containing variance due to cognitive abilities and self efficacy beliefs. Thus, the need to guard against faking seems to be unnecessary.

### 5.3. Summary and outlook

The present project aimed at exploring the impact of situational demand on the construct validity and the criterion validity of a personality questionnaire. It was shown that situational demand causes correlations between personality factors and influences the variance composition of a trait. While construct validity is affected by situational demand, the results also show that the criterion validity is only indirectly

affected. This means that the explained variance does not change due to situational demand. However, only facets contribute to the explanation and the ranking in the personality factors assessed also changes. Thus, the impact of situational demand on personality factors was confirmed.

Another goal of the project was the exploration of the psychological process behind faking. Based on the different results, a model of responding to situational demand was proposed.

Finally, a new method for correcting faked personality scores was tested, but did not prove useful with only six items.

Future research should try to find more evidence for the proposed model. So far, variance due to situational demand could not be explained totally. Actually, only a small part of the variance could be accounted for. It can be assumed that specific knowledge is one of the main sources of individual differences in faking. Other sources should be investigated since they might allow a better understanding of the psychological process as well as the criterion validity of personality questionnaires. Analyzing faked and honest personality facets in regard to their predictive power might also foster the knowledge of which specific personality attributes contribute to better performance outcomes.

This project was successful in exploring and explaining situational influence on personality questionnaires. However, it was only a first step and many more interesting research questions have been proposed. Hopefully, other researchers will take up the ideas presented here and join in the quest of finding out how state and trait interact in influencing people's behavior and performance.

## References

- Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence, 22*, 227-257.
- Allik, J. & McCrae, R. R. (2004). Escapable Conclusions: Toomela (2003) and the Universality of Trait Structure. *Journal of Personality and Social Psychology, 87*, 261-265.
- Allport, G. W. (1961). *Pattern and Growth in Personality*. New York: Holt, Rinehart & Winston.
- Allport, G. W. & Odbert, H. S. (1936). *Trait names: A psycho-lexical study. Psychological Monographs, 47* (1, Whole No. 2 1 1).
- Aluja, A., Garcia, O., Garcia, L. F. & Seisdedos, N. (2005). Invariance of the "NEO-PI-R" factor structure across exploratory and confirmatory factor analyses. *Personality and Individual Differences, 38*, 1879.
- Amthauer, R., Brocke, B., Liepmann, D. & Beauducel, A. (2001). *I-S-T 2000 R; Intelligenz-Struktur-Test 2000 R; Intelligence Structure Test 2000*. Hogrefe.
- Asendorpf, J. B. (2006). On artifacts and meaning in person-centred analyses: Comment on McCrae, Terracciano, Costa and Ozer (2005). *European Journal of Personality, 20*, 45-47.
- Ashton, M.-C., Lee, K., Perugini, M., Szarota, P., de-Vries, R.-E., Di-Blas, L., 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.

- Austin, E. J., Deary, I. J. & Egan, V. (2006). Individual differences in response scale use: Mixed Rasch modelling of responses to NEO-FFI items. *Personality And Individual Differences, 40*, 1235-1245.
- Baer, R. A., Wetter, M. W., Nichols, D. S., Greene, R. & et al. (1995). Sensitivity of MMPI-2 validity scales to underreporting of symptoms. *Psychological Assessment, 7*, 419-423.
- Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology, 69*, 49-56.
- Barrick, M. R. & Mount, M. K. (1991). The BIG Five Personality Dimensions and Job Performance: A Meta-Analysis. *Personnel Psychology, 44*, 1-26.
- 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. Vol, 9*, 9-30.
- Beauducel, A., Brocke, B. & Liepmann, D. (2001). Perspectives on fluid and crystallized intelligence: facets for verbal, numerical, and figural intelligence. *Personality and Individual Differences, 30*, 977-994.
- Beauducel, A. & Wittmann, W. W. (2005). Simulation Study on Fit Indexes in CFA Based on Data With Slightly Distorted Simple Structure. *Structural Equation Modeling, 41-75*.
- 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, 845-876.
- Blackburn, R., Logan, C., Renwick, S. J. D. & Donnelly, J. P. (2005). Higher-order dimensions of personality disorder: Hierarchical structure and relationships with the five-factor model, the interpersonal circle, and psychopathy. *Journal of Personality Disorders*, 19, 597-623.
- Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin*, 117, 187-215.
- Bollen, E. & Wojciechowski, F. L. (2004). Anorexia nervosa subtypes and the big five personality factors. *European Eating Disorders Review*, 12, 117-121.
- Bono, J. E. & Judge, T. A. (2004). Personality and transformational and transactional leadership: A meta-analysis. *Journal Of Applied Psychology*, 89, 901-910.
- Bouchard, T. J. (1997). The Genetics of Personality, In Blum, K. & Noble, E. P. (eds.), *Handbook of Psychiatric Genetics* (pp. 273-296) Boca Raton, FL: Circle Press Inc.
- Bowen, C. C., Martin, B. A. & Hunt, S. T. (2002). A comparison of ipsative and normative approaches for ability to control faking in personality questionnaires. *International Journal of Organizational Analysis*, 10, 240-259.
- Brand, C., Egan, V. & Deary, I. J. (1993). Personality and general intelligence, (1993). Bonaiuto, P., Van Heck, G. L. (Eds), *Personality psychology in Europe, Vol. 4.* (pp.203 228). Tilburg, Netherlands: Tilburg University Press. xiv, 416 pp.

- Buehner, M. (2006). *Introduction to Test- and Questionnaire Construction [Einfuehrung in die Test- und Fragebogenkonstruktion]*. Pearson.
- Buehner, M., Krumm, S. & Pick, M. (2005). Reasoning is Working Memory and not Attention. *Intelligence, 33*, 251-272.
- Buehner, M., Ziegler, M., Krumm, S. & Schmidt Atzert, L. (2006). Ist der I-S-T 2000 R Rasch-skalierbar? [Is the I-S-T 2000 R Rasch scaleable?]. *Diagnostica, 92*, 119-130.
- Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Lawrence Erlbaum Associates.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Cattell, R. B. (1943a). The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology, 38*, 476-506.
- Cattell, R. B. (1943b). The description of personality: I. Foundations of trait measurement. *Psychological Review, 50*, 559-594.
- 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*, 694-704.
- Chamorro Premuzic, T. & Furnham, A. (2003). Personality traits and academic examination performance. *European Journal of Personality, 17*, 237-250.

- Chamorro Premuzic, T. & Furnham, A. (2004). A possible model for understanding the personality-intelligence interface. *British Journal of Psychology, 95*, 249-264.
- Chamorro Premuzic, T., Furnham, A. & Moutafi, J. (2004). The relationship between estimated and psychometric personality and intelligence scores. *Journal of Research in Personality, 38*, 505-513.
- Cheung, G. W. & Rensvold, R. B. (2001). The effects of model parsimony and sampling error on the fit of structural equation models. *Organizational Research Methods, 4*, 236-264.
- Christiansen, N. D., Burns, G. N. & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*, 267-307.
- Collins, J.-M. & Gleaves, D.-H. (1998). Race, job applicants, and the Five-Factor Model of Personality: Implications for Black psychology, industrial/organizational psychology, and the Five-Factor Theory. *Journal of Applied Psychology, 83*, 531-544.
- Costa, P. T., Jr. & McCrae, R. R. (1992a). Reply to Eysenck. *Personality and Individual Differences, 13*, 861-865.
- Costa, P. T., Jr. & McCrae, R. R. (2006). Age Changes in Personality and Their Origins: Comment on Roberts, Walton, and Viechtbauer (2006). *Psychological Bulletin, 132*, 26-28.
- Costa, P. T. & McCrae, R. R. (1980). Still stable after all these years: Personality as a key to some issues in adulthood and old age, In Baltes, P. B. & Brim, O.G.

- (Eds.), *Life Span development und behaviors* (pp. 65-102). New York: Academic Press.
- Costa, P. T. & McCrae, R. R. (1992b). *Revised NEO personality inventory (NEO-PI-R) and NEO Five-Factor inventory (NEO-FFI): Professional Manual*. Odessa: Psychological Assessment Resources.
- Costa, P. T. & McCrae, R. R. (1995). Solid ground in the wetlands of personality: A reply to Block. *Psychological Bulletin*, *117*, 216-220.
- Crowne, D. & Marlowe, D. (1960). A new scale of social desirability of psychopathology. *Journal of Consulting Psychology*, *24*, 329-354.
- Dacosta, M. & Halmi, K. A. (1992). Classifications of Anorexia-Nervosa - Question of Subtypes. *International Journal of Eating Disorders*, *11*, 305-313.
- De Fruyt, F., Aluja, A., Garcia, L. F., Rolland, J.-P. & Jung, S. C. (2006). Positive Presentation Management and Intelligence and the Personality Differentiation by Intelligence Hypothesis in Job Applicants. *International Journal of Selection and Assessment*, *14*, 101-112.
- De Fruyt, F., Bockstaele, M., Taris, R. & Van Hiel, A. (in press). Police Interview Competencies: Assessment and Associated Traits. *European Journal of Personality*,
- Deary, I. J. & Pagliari, C. (1991). The strength of g at different levels of ability: Have Detterman and Daniel rediscovered Spearman's "law of diminishing returns"? *Intelligence*, *15*, 247-250.

- Deinzer, R., Steyer, R., Eid, M., Notz, P., Schwenkmezger, P., Ostendorf, F. & Neubauer, A. (1995). Situational effects in trait assessment: The FPI, NEOFFI, and EPI questionnaires. *European Journal of Personality*. Vol, 9, 1-23.
- Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, 73, 1246-1256.
- Digman, J. M. & Shmelyov, A. G. (1996). The structure of temperament and personality in Russian children. *Journal of Personality and Social Psychology*, 71, 341-351.
- Dillman, D. A. & Redline, C. D. (2004). Testing Paper Self-Administered Questionnaires: Cognitive Interview and Field Test Comparisons, In Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J. & Singer, E. (Eds), *Methods for Testing and Evaluating Survey Questionnaires* Hoboken: John Wiley & Sons, Inc.
- 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, 122-133.
- Eysenck, H. J. (1992). A reply to Costa and McCrae: P or A and C--the role of theory. *Personality and Individual Differences*, 13, 867-868.
- Eysenck, H. J. & Eysenck, S. B. G. (1964). *The Eysenck Personality Inventory*. University of London Press, London.
- Fahrenberg, J., Hampel, R. & Selg, H. (1984). Das Freiburger Persönlichkeitsinventar (FPI und FPI-R), Handbuch [The Freiburg Personality Inventory (FPI and FPI-R), Manual].

- Farsides, T. & Woodfield, R. (2003). Individual differences and undergraduate academic success: the roles of personality, intelligence, and application. *Personality and Individual Differences, 34*, 1225-1243.
- Ferrando, P. J. & Chico, E. (2001). Detecting dissimulation in personality test scores: A comparison between person-fit indices and detection scales. *Educational and Psychological Measurement, 61*, 997-1012.
- Fowler, F. J. (1995). *Improving survey questions. Design and evaluation. Applied Social Research Methods Series (Chapter 5)*. Sage.
- Freud, S. (1910). The origin and development of psychoanalysis. *American Journal of Psychology, 21*, 181-218.
- Furnham, A. (2002). Dissimulation, intelligence and personality. *Social Behavior and Personality, 30*, 527-532.
- Furnham, A. & Chamorro-Premuzic, T. (2004). Personality and intelligence as predictors of statistics examination grades. *Personality and Individual Differences, 37*, 943-955.
- Furnham, A., Chamorro Premuzic, T. & McDougall, F. (2002). Personality, cognitive ability, and beliefs about intelligence as predictors of academic performance. *Learning and Individual Differences, 14*, 47-64.
- Furnham, A. & Crump, J. (2005). Personality traits, types, and disorders: An examination of the relationship between three self-report measures. *European Journal of Personality, 19*, 167-184.

- Furnham, A., Moutafi, J. & Chamorro Premuzic, T. (2005). Personality and Intelligence: Gender, the Big Five, Self-Estimated and Psychometric Intelligence. *International Journal of Selection and Assessment*, 13, 11-24.
- Furnham, A., Moutafi, J. & Paltiel, L. (2005). Intelligence in Relation to Jung's Personality Types. *Individual Differences Research*, 3, 2-13.
- Furnham, A. F. (1997). Knowing and faking one's Five-Factor personality score. *Journal of Personality Assessment*, 69, 229-243.
- Goldberg, L. R. (1981). Language and individual differences: the search for universals in personality lexicons. *Review of Personality and Social Psychology*, 2, 141-165.
- Goldberg, L. R. (1992). The development of marker variables for the Big-Five factor structure. *Psychological Assessment*, 4, 26-42.
- Gurrera, R. J., Nestor, P. G., O'Donnell, B. F., Rosenberg, V. & McCarley, R. W. (2005). Personality differences in schizophrenia are related to performance on neuropsychological tasks. *Journal of Nervous and Mental Disease*, 193, 714-721.
- Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and mimic approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, 66, 373-388.
- Hancock, G. R. & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems., Cudeck, R., Toit, S. du & Sörbom, D., *Structural Equation*

- Modeling: Present and Future - Festschrift in honor of Karl Jöreskog* (195-216). Lincolnwood, IL: Scientific Software International, Inc.
- Harris, J. A., Vernon, P. A. & Jang, K. L. (2005). Testing the differentiation of personality by intelligence hypothesis. *Personality and Individual Differences*, 38, 277-286.
- Hough, L. M. (1998). Effects of intentional distortion in personality measurement and evaluation of suggested palliatives. *Human Performance*, 11, 209-244.
- Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D. & et al. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology*, 75, 581-595.
- Howell, D. C. (2002). *Statistical Methods for Psychology*. Duxbury: Thomson Learning.
- Hu, L. t. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.
- Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, 96, 72-98.
- Hunter, J. E. & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks: Sage Publications, Inc.
- John, O. P., Angleitner, A. & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality*, 2, 171-203.

- Judge, T. A. & Ilies, R. (2002). Relationship of personality to performance motivation: A meta-analytic review. *Journal of Applied Psychology, 87*, 797-807.
- Kallasmaa, T., Allik, J., Realo, A. & McCrae, R. R. (2000). The Estonian version of the NEO-PI-R: An examination of universal and culture-specific aspects of the five-factor model. *European Journal of Personality, 14*, 265-278.
- Konstabel, K., Aavik, T. & Allik, J. (in press). Social Desirability and Consensual Validity of Personality Traits. *European Journal of Personality*,
- Kuncel, N. R., Hezlett, S. A. & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127*, 162-181.
- Kuncel, N. R., Hezlett, S. A. & Ones, D. S. (2004). Academic Performance, Career Potential, Creativity, and Job Performance: Can One Construct Predict Them All? *Journal of Personality and Social Psychology, 86*, 148-161.
- LaHuis, D. M., Martin, N. R. & Avis, J. M. (2005). Investigating Nonlinear Conscientiousness-Job Performance Relations for Clerical Employees. *Human Performance, 18*, 199-212.
- Liebert, R. M. & Liebert, L. L. (1998). *Personality Strategies & Issues*. Pacific Grove: Brooks/Cole Publishing.
- Lounsbury, J. W., Sundstrom, E., Loveland, J. L. & Gibson, L. W. (2002). Broad versus narrow personality traits in predicting academic performance of adolescents. *Learning and Individual Differences, 14*, 65-75.

- Marcus, B., Machilek, F. & Schutz, A. (2006). Personality in cyberspace: Personal Web sites as media for personality expressions and impressions. *Journal of Personality and Social Psychology, 90*, 1014-1031.
- Marsh, H. W., Hau, K. T. & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling, 11*, 320-341.
- Martin, B. A., Bowen, C. C. & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences, 32*, 247-256.
- McClelland, D. C. (1987). *Human Motivation*. New York: Cambridge University Press.
- McClelland, D. C., Koestner, R. & Weinberger, J. (1989). How do self-attributed and implicit motives differ? *Psychological Review, 96*, 690-702.
- McCrae, R. R. (1992). Editor's introduction to Tupes and Christal. *Journal of Personality, 60*, 217-219.
- McCrae, R. R., Terracciano, A., Costa, P. T. & Ozer, D. J. (2006a). From types to typological thinking: A reply to Asendorpf. *European Journal of Personality, 20*, 49-51.
- McCrae, R. R., Terracciano, A., Costa, P. T. & Ozer, D. J. (2006b). Person-factors in the California adult Q-set: Closing the door on personality trait types? *European Journal of Personality, 20*, 29-44.

- McCrae, R. R., Zonderman, A. B., Bond, M. H., Costa, P. T. & Paunonen, 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.
- McFarland, L. A. & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*, 812-821.
- Mielke, R. (1990). Ein Fragebogen zur Wirksamkeit der Selbstdarstellung [Efficacy of self-presentation questionnaire]. *Zeitschrift für Sozialpsychologie [Journal of Social Psychology], 21*, 162-170.
- Mischel, W. (1968). *Personality and assessment*. John Wiley & Sons, Inc.
- Mischel, W. & Peake, P. K. (1982). Beyond deja vu in the search for cross-situational consistency. *Psychological Review, 89*, 730-755.
- Möttus, R., Allik, J. & Pullmann, H. (2006). Does personality vary across ability levels? A study using self and other ratings. *Journal of Research in Personality*, (in press).
- Mount, M. K. & Barrick, M. R. (1998). Five reasons why the "big five" article has been frequently cited. *Personnel Psychology, 51*, 849-857.
- Mount, M. K., Barrick, M. R., Scullen, S. M. & Rounds, J. (2005). Higher-order dimensions of the big five personality traits and the big six vocational interest types. *Personnel Psychology, 58*, 447-478.

- Moutafi, J., Furnham, A. & Crump, J. (2003). Demographic and personality predictors of intelligence: A study using the Neo personality inventory and the Myers-Briggs type indicator. *European Journal of Personality, 17*, 79-94.
- Moutafi, J., Furnham, A. & Paltiel, L. (2004). Why is conscientiousness negatively correlated with intelligence? *Personality and Individual Differences, 37*, 1013-1022.
- Moxnes, P. A. (1996). On the deeper meaning of the Big 5 and personality disorders. *International Journal of Psychology, 31*, 2804-2804.
- Murphy, K. R. & Davidshofer, C. O. (2001). *Psychological Testing*. New Jersey: Prentice Hall.
- Nguyen, N. T., Biderman, M. D. & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment, 13*, 250-260.
- Norman, W. T. (1963). Toward an Adequate Taxonomy of Personality Attributes - Replicated Factor Structure in Peer Nomination Personality Ratings. *Journal of Abnormal Psychology, 66*, 574-583.
- Norman, W. T. (1967). *2800 personality trait descriptors: Normative operating characteristics for a university population*. Ann Arbor: University of Michigan, Department of Psychological Sciences.
- 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*, 660-679.

- Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct, Braun, H. I., Jackson, D. N. & Wiley, D. E., *The role of constructs in psychological and educational measurement*. (49-69). Mahwah, NJ, US: Lawrence Erlbaum Associates.
- Paulhus, D. L., Harms, P. D., Bruce, M. N. & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology*, *84*, 890-904.
- Pauls, C.-A. & Crost, N.-W. (2005a). Cognitive Ability and Self-Reported Efficacy of Self-Presentation Predict Faking on Personality Measures. *Journal of Individual Differences*, *26*, 194-206.
- Pauls, C. A. & Crost, N. W. (2004). Effects of faking on self-deception and impression management scales. *Personality and Individual Differences*, *37*, 1137-1151.
- Pauls, C. A. & Crost, N. W. (2005b). Effects of different instructional sets on the construct validity of the NEO-PI-R. *Personality and Individual Differences*, *39*, 297-308.
- Peeters, H. & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement*, *65*, 70-89.
- Peeters, M. A. G., Rutte, C. G., van Tuijl, H. E. J. M. & Reymen, I. M. M. J. (2006). The Big Five personality traits and individual satisfaction with the team. *Small Group Research*, *37*, 187-211.

- Pekrun, R., Goetz, T., Perry, R. P., Kramer, K., Hochstadt, M. & Molfenter, S. (2004). Beyond test anxiety: Development and validation of the Test Emotions Questionnaire (TEQ). *Anxiety Stress and Coping, 17*, 287-316.
- Pekrun, R., Goetz, T., Titz, W. & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist, 37*, 91-105.
- Picardi, A., Toni, A. & Caroppo, E. (2005). Stability of Alexithymia and its relationships with the 'big five' factors, temperament, character, and attachment style. *Psychotherapy and Psychosomatics, 74*, 371-378.
- Pulver, A. (1999). Vaimsete võimete skaala VVS2. Unpublished manual. Tartu, Estonia: University of Tartu, Department of Psychology.
- Rindermann, H. & Neubauer, A. C. (2001). The influence of personality on three aspects of cognitive performance: Processing speed, intelligence and school performance. *Personality and Individual Differences, 30*, 829-842.
- 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*, 1-25.
- Robertson, I. T., Baron, H., Gibbons, P., MacIver, R. & Nyfield, G. (2000). Conscientiousness and managerial performance. *Journal of Occupational and Organizational Psychology, 73*, 171-180.
- Rogers, R. (1997). Researching Dissimulation, In Rogers, R., *Clinical assessment of malingering and deception* (309-327). New York: Guilford.

- 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*, 634-644.
- Rost, J. (1985). A Latent Class Model for Rating Data. *Psychometrika, 50*, 37-49.
- Rost, J. (1991). Alogistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology, 44*, 75-92.
- Rost, J., Carstensen, C. H. & Von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires, In Rost, J. & Langeheine, R. (Eds.), *Applications of latent trait and latent class models in the social sciences* Waxmann.
- Rost, J., Carstensen, C. H. & von Davier, M. (1999). Are the Big Five Rasch scalable? A reanalysis of the NEO-FFI norm data. *Diagnostica, 45*, 119-127.
- Salgado, J. (2002). The Big Five personality dimensions and counterproductive behaviors. *International Journal of Selection and Assessment, 10*, 117-125.
- Salgado, J. F. (1997). The five factor model of personality and job performance in the European Community. *Journal of Applied Psychology, 82*, 30-43.
- Salgado, J. F. (2003). Predicting job performance using FFM and non-FFM personality measures. *Journal of Occupational and Organizational Psychology, 76*, 323-346.

- Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C. & de Fruyt, F. (2003). International Validity Generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology, 56*, 573-605.
- Saucier, G. (2003). Factor structure of English-language personality type-nouns. *Journal of Personality and Social Psychology, 85*, 695-708.
- Saucier, G., Georgiades, S., Tsaousis, I. & Goldberg, L.-R. (2005). The Factor Structure of Greek Personality Adjectives. *Journal of Personality and Social Psychology, 88*, 856-875.
- Saulsman, L. M. & Page, A. C. (2003). Can trait measures diagnose personality disorders? *Current Opinion in Psychiatry, 16*, 83-88.
- Saulsman, L. M. & Page, A. C. (2004). The five-factor model and personality disorder empirical literature: A meta-analytic review. *Clinical Psychology Review, 23*, 1055-1085.
- Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.
- 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*, 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*, 613-621.

- Schmitz, N., Kugler, J. & Rollnik, J. (2003). On the relation between neuroticism, self-esteem, and depression: Results from the national comorbidity survey. *Comprehensive Psychiatry*, *44*, 169-176.
- Shao, L. & Webber, S. (2006). A cross-cultural test of the 'five-factor model of personality and transformational leadership'. *Journal of Business Research*, *59*, 936-944.
- Sifneos, P. E. (1973). Prevalence of Alexithymic Characteristics in Psychosomatic Patients. *Psychotherapy and Psychosomatics*, *22*, 255-262.
- Smith, D. B. & Ellingson, J. E. (2002). Substance versus style: A new look at social desirability in motivating contexts. *Journal of Applied Psychology*, *87*, 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*, 304-315.
- Smits, D. J. M. & Boeck, P. D. (2006). From BIS/BAS to the Big Five. *European Journal of Personality*, *20*, 255-270.
- Sokolowski, K., Schmalt, H. D., Langens, T. A. & Puca, R. M. (2000). Assessing achievement, affiliation, and power motives all at once: The Multi-Motive Grid (MMG). *Journal of Personality Assessment*, *74*, 126-145.
- Spearman, C. (1927). *The abilities of man*. Oxford, England: Macmillan.

- 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*, 943-953.
- Steyer, R. (2005). Analyzing Individual and Average Causal Effects via Structural Equation Models. *Methodology: -European-Journal-of-Research-Methods-for-the-Behavioral-and-Social-Sciences, 1*, 39-54.
- Steyer, R., Ferring, D. & Schmitt, M. J. (1992). States and traits in psychological assessment. *European-Journal-of-Psychological-Assessment, 8*, 79-98.
- Steyer, R., Schmitt, M. & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European-Journal-of-Personality, Vol 13*, 389-408.
- Taylor, G. J. & Bagby, R. M. (2004). New trends in alexithymia research. *Psychotherapy and Psychosomatics, 73*, 68-77.
- ter Laak, J. J. F., DeGoede, M. P. M. & Brugman, G. M. (2001). Teachers' judgments of pupils: Agreement and accuracy. *Social Behavior and Personality, 29*, 257-270.
- Toomela, A. (2003). Relationships between personality structure, structure of word meaning, and cognitive ability: A study of cultural mechanisms of personality. *Journal of Personality and Social Psychology, 85*, 723-735.
- Topping, G. D. & O'Gorman, J. G. (1997). Effects of faking set on validity of the NEO-FFI. *Personality and Individual Differences, 23*, 117-124.

- 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*, 306-322.
- Viswesvaran, C. & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197-210.
- Viswesvaran, C., Ones, D. S. & Hough, L. M. (2001). Do impression management scales in personality inventories predict managerial job performance ratings? *International Journal of Selection and Assessment, 9*, 277-289.
- Wagner-Menghin, M. (1998). *Lexikon-Wissen-Test [Lexical Knowledge Test]*. Schuhfried.
- Willis, G. (2004). Cognitive Interviewing Revisited: A Useful Technique, in Theory? In Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J. & Singer, E. (Eds), *Methods for Testing and Evaluating Survey Questionnaires* (299-317). John Wiley & Sons, Inc.
- Zeidner, M. & Matthews, G. (2000). Intelligence and personality, (2000). Sternberg, Robert J (Ed), *Handbook of intelligence*. (pp.581 610). New York, NY, US: Cambridge University Press. xiii, 677 pp.
- Zickar, M.-J. & Robie, C. (1999a). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology, 84*, 551-563.
- Zickar, M. J. & Burnfield, J. L. (2003). A Monte Carlo investigation of the polytomous mixed Rasch model. *Unpublished manuscript*,

- 
- 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*, 168-190.
- Zickar, M. J. & Robie, C. (1999b). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology, 84*, 551-563.
- Ziegler, M. & Bühner, M. (submitted). Personality Structure: A Question of Trait and State Interaction.
- Ziegler, M., Bühner, M., Krumm, S. & Schmidt Atzert, L. (submitted-a). Motivated, Unmotivated, or Faking? Susceptibility of Three Achievement Motivation Tests Measures to Faking: Questionnaire, Semi-projective, and Objective.
- Ziegler, M., Bühner, M. & Toomela, A. (submitted-b). A Reanalysis of Toomela (2003): Personality Structure – A Question of IQ or Situational Demand?

# MATTHIAS ZIEGLER

## PERSÖNLICHE INFORMATIONEN

---

Familienstand: ledig  
Nationalität: deutsch  
Alter: 28  
Geburtsort: Heiligenstadt  
Eltern: Petra und Eberhard Ziegler

## SPEZIELLE QUALIFIKATIONEN

---

DIN 33430 Volllizenz für Auftragnehmer (A-Lizenz)  
Mitglied der Technical Advisor Group on Faking des Educational Testing Service

## AUSBILDUNG

---

1991 – 1998 Ernst Abbe Gymnasium Eisenach Abitur (1,0)  
1999 – 2005 Philipps-Universität Marburg Diplom in Psychologie („Mit Auszeichnung“)  
Zusatzfach Betriebswirtschaftslehre mit der Abschlussnote „gut“  
2005 - 2007 Promotion bei Prof. Dr. Markus Bühner, Psychologische Methodenlehre und Evaluation, Department Psychologie, LMU mit dem Thema „Persönlichkeit und Intelligenz in Auswahl-situationen“ („Summa cum laude“)

## Berufstätigkeit

01.04.2005 – 31.03.2006: Wissenschaftlicher Projektmitarbeiter in einem durch die Deutsche Rentenversicherung Bund finanzierten Projekt  
01.04.2006 – 30.09.2006: Wissenschaftlicher Mitarbeiter am Department Psychologie der LMU München, Abteilung Pädagogische Psychologie und Diagnostik  
01.10.2006 – 31.03.2007: Wissenschaftlicher Mitarbeiter am Department Psychologie der LMU München, Abteilung Psychologische Methodenlehre und Evaluation

## WEITERE BERUFSTÄTIGKEITEN

---

seit 2003 freier Mitarbeiter bei SHL Deutschland mit den Schwerpunkten:  
Konzeption von Einzel- und Gruppenassessments  
Durchführung von Einzel- und Gruppenassessments  
Verhaltensbeobachtung  
Beobachtercoaching  
Gutachtenerstellung  
Führen von Feedbackgesprächen  
Erstellen eines Konzeptes zur Lehrlingsauswahl für WINCANTON Trans European (Deutschland) GmbH, Zweigniederlassung Eisenach  
seit 2005 freier Mitarbeiter bei der Deutschen Flugsicherung  
Beratung für die Firma Organon

## WEITERE TÄTIGKEITEN

- 
- Mitgründer und 2004 Präsident des Debattierclubs „Hartshornes House“ in Marburg
- 1996 – 2004 Nachhilfelehrer in der Schülerhilfe Eisenach für die Fächer Mathematik und Englisch
- 2002 – 2003 Nachhilfelehrer für Chemie in einem Wohnprojekt der Deutschen Lebenshilfe e.V. in Eisenach
- Seit 01.04.2006 Consultant beim TestLab der LMU
- 2005/2006 Durchführung von Workshops zur DIN 33430

## PUBLIKATIONEN

- 
- Bühner, M. & Mangels, M., Krumm, S. & Ziegler, M. (2005). Are working memory and attention related constructs. *Journal of Individual Differences*, 26 (3), 121-131.
- Bühner, M., Ziegler, M., Krumm, S. & Schmidt-Atzert, L. (2006). Ist der I-S-T 2000 R Rasch-skalierbar? *Diagnostica*, 52 (3).
- Bühner, M., Krumm, S., Ziegler, M., & Pluecken, T. (2006). Cognitive abilities and their interplay: Reasoning, crystallized intelligence, working memory components and sustained attention. *Journal of Individual Differences*.
- Bühner, M., Ziegler, M., Bohnes, B., & Lauterbach, K. (2006). Übungseffekte in der Testbatterie zur Aufmerksamkeitsprüfung und im Test d2 bei wiederholter Durchführung. *Zeitschrift für Neuropsychologie* (3).
- Ziegler, M., Schmidt-Atzert, L., Bühner, M., & Krumm, S. (submitted). Faking Susceptibility of Different Measurement Methods: Questionnaire, Semi-Projective, and Objective.
- Krumm, S., Ziegler, M. & Bühner, M., (submitted). Intelligence and working memory as predictors of school performance.
- Müller, J., Bühner, M., Ziegler, M. & Sahin, L. (submitted). Are Alexithymia, Ambivalence over Emotional Expression, and Social Insecurity overlapping constructs?
- Krumm, S., Schmidt-Atzert, L., Bühner, M., Ziegler, M., Michalczyk, K., & Arrow, K., (submitted). Working Memory is not necessary for the Prediction of Fluid Intelligence! Towards a Parsimonious Model of Cognitive Abilities.

## VORTRÄGE

- 
- Der Zusammenhang zwischen Persönlichkeit und Intelligenz in Auswahl-situationen. Doktorandenworkshop der Fachgruppe für Differentielle Psychologie, Persönlichkeitspsychologie und Psychologische Diagnostik in Landau am 08. September 2005 (Ziegler, M.)
- Unmotivierte, Hochmotivierte oder Simulanten? Ein Vergleich der Simulationsresistenz verschiedener Leistungsmotivationsinventare. 8. Arbeitstagung Differentielle Psychologie, Persönlichkeitspsychologie und Psychologische Diagnostik. 28. und 29.09.2005 in Marburg (Ziegler, M. Schmidt - Atzert, L., Bühner, M. & Krumm, S.).
- Faking in Persönlichkeitstests. Deutsche Gesellschaft für Personalwesen am 09. Januar 2006 in Hannover. (Ziegler, M.).

Einsatz der kognitiven Survey Technik zur Verbesserung des Fragebogens zur Patientenzufriedenheit in der Rehabilitation. 15. Rehabilitationswissenschaftliches Kolloquium vom 13. bis 15. März 2006 in Bayreuth (Ziegler, M., Bühner, M., & Raspe, H.).

The influence of situational demands on the structure of the big five. 13th European Conference on Personality vom 22. bis 26. Juli in Athen (Ziegler, M. & Bühner, M.).

Big-Five und die Vorhersage von Ausbildungserfolg: Persönlichkeit oder Faking? 45. Kongress der DGPs 2006 vom 17. - 21. September 2006 in Nürnberg (Ziegler, M. & Bühner, M.)

People fake! - So what? ETS Miniconference on Faking, 13.-14.10.2006, Princeton, USA (Ziegler, M.).

#### PRAKTIKA

---

2002 Change- und Projektmanagement bei der Dresdner Bank AG

2003 Forschungspraktikum in der AG Diagnostik an der Philipps-Universität in Marburg

2003 SHL Deutschland GmbH

#### STIPENDIEN

---

2002 – 2005 Stipendiat bei e-fellows Deutschland