universität wien

Dissertation

Titel der Dissertation

Context effects in personnel selection: The influence of different contexts on the self-description in personality questionnaires as well as on test scores of cognitive ability tests and objective personality tests

Verfasserin

Mag. rer. nat. Lale Khorramdel

angestrebter akademischer Grad

Doktorgrad der Naturwissenschaften

Wien, im Juni 2010

Studienkennzahl lt. Studienblatt:	A 091 298
Dissertationsgebiet lt. Studienblatt:	Psychologie
Betreuer:	Univ.-Prof. Dr. Klaus D. Kubinger

In liebevoller Erinnerung an meine Großeltern

Johanna und Simon Kienleitner

Acknowledgments

First and foremost I offer my sincerest gratitude to my supervisor and mentor, Univ.-Prof. Dr. Klaus D. Kubinger, who has supported me throughout my thesis with his patience, knowledge and unflinching encouragement in various ways. His truly scientist intuition and passion exceptionally inspired and enriched my growth as a student and a scientist. I am indebted to him more than he knows.

I gratefully acknowledge the Centre of Testing and Consulting[1] for providing the necessary data from projects related to personnel selection, which made the department a backbone of this research and so to this thesis.

I convey special acknowledgement to my colleagues Martina Frebort, Christine Hohensinn, Stefana Holocher-Ertl, and Philipp Sonnleitner for their outstanding collegiality and friendship over the years. Martina's contribution and involvement as my co-author in two of the five articles and her professional support triggered and nourished the quality of the thesis.

Many thanks go to my department colleagues Sandra Hofer, Simon Lehner, Manuel Reif, Eva Schleicher, Lisbeth Weitensfelder, and Takuya Yanagida for giving me such a pleasant time when working together with them and for creating such a great friendship at the office.

Additional thanks go to Alexander Uitz for his assistance with the data collection and to Georg Wilflinger for pre-reviewing one of the articles.

Words fail me to express my appreciation to my family for their support and persistent confidence in me.

[1] Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna; www.testzentrum.at

Table of Contents

Abstract – English 1
Abstract – Deutsch 3

1. General Introduction **5**
1.1. Faking 6
1.1.1. Faking good, faking bad: a definition of the common terminology 6
1.1.2. Faking good on personality questionnaires in personnel selection 7
1.1.3. Strategies to deal with faking 9
1.2. Context effects 12

2. Contribution of the current doctoral thesis **15**
2.1. Methodical approach 15
2.2. The effect of speediness (time limit), response format and warning instruction on impression management in personality questionnaires 15
2.2.1. Paper 1 and Paper 4 15
2.2.2. Additional Study 17
2.2.2.1. Results (Additional Study) 17
2.2.3. Discussion and critical reflection (Paper 1, Paper 4, Additional Study) 20
2.3. Questionnaire length and impression management 24
2.3.1. Paper 2 24
2.3.2. Discussion and critical reflection (Paper 2) 25
2.4. The effect of test order on test performance (with special regard to objective personality tests) 26
2.4.1. Paper 3 and Paper 5 26
2.4.2. Discussion and critical reflection (Paper 3, Paper 5) 28

3. General Discussion and Prospects **29**

4. References **32**

5. Original Papers of the Doctoral Thesis (Paper 1 – 3) **41**
5.1. Paper 1 43
5.2. Paper 2 65
5.3. Paper 3 99

6. Tables and Additional Papers (Appendix 1 – 2) **123**
6.1. Appendix 1 (Table 4, Paper 1; Table 1, Additional Study) 125
6.2. Appendix 2 (Paper 4 – 5) 131

7. Curriculum Vitae **183**

8. Publications **184**

Abstract - English

Responses on psychological tests or questionnaires are not always determined by mapping the trait directly on a response scale as they can also be influenced by numerous other variables (person, test, and situation related) resulting in context effects, which might sometimes harm the measurement. Intentional response distortion (impression management, faking good, socially desirable response behaviour) in personality questionnaires and the effects of different test orders can both be described as such contexts effects. The current thesis consists of three scientific articles (papers) and an additional study, which investigate the effects of different strategies (such as different response formats, instructions, limited response time, and questionnaire length) to decrease socially desirable response behaviour, as well as the effects of different test orders on the test performance in objective personality tests and cognitive ability tests, by examining applicants. Their findings are completed by two further scientific articles (papers), which investigate the effects of different response formats on impression management and the effects of different test orders on test performance once again, however with different samples and partially different questionnaires and tests. Results provide evidence for the hypothesis that response scales with a higher number of response alternatives (analogue scales or 6-point rating scales), instructions with warnings (that fakers can be detected), and items positioned at the end of a questionnaire lead to less socially desirable responses than response scales with only two response alternatives (dichotomous response formats or 2-point rating scales), instructions without a warning, and items positioned at the beginning of a questionnaire. These strategies seem to make it more difficult for test-takers to fit their responses to a faking schema (Holden & Hibbs, 1995; Holden, Kroner, Fekken, & Popham, 1992) or to influence test-takers' motivation to fake (Rothstein & Goffin, 2006). A limited response time was shown to lead to less socially desirable responses as well, but only in combination with an analogue scale or a warning instruction. The effects of the different strategies were either not consistent within or across the different studies, which leads to the assumption that there is an interaction between person related variables and the content of the questionnaire scales or item wording. Moreover, different test orders influenced the test performance in objective personality tests but not in cognitive ability tests. Test-takers, who worked on cognitive ability tests first and on objective personality tests second, were shown to be more decisive and to have a lower tolerance to frustration in the objective personality tests than test-takers, who had worked on objective personality tests first. Again, results were not consistent across the different studies. Altogether, evidence is provided for the theory that faking is a variable of individual differences (Mueller-Hanson, Heggestad, & Thornton, 2006; Viswesvaran, & Ones, 1999), and for the theory that context effects depend on differences in test-takers' motivation and cognitive ability (Schwarz, Hippler, & Noelle-Neumann, 1992).

Abstract - Deutsch

Das Antwortverhalten in psychologisch-diagnostischen Verfahren wird nicht ausschließlich durch die entsprechenden Eigenschaften oder Fähigkeiten bestimmt, sondern von zahlreichen anderen (personen-, verfahrens- und situationsspezifischen) Variablen beeinflusst. Diese resultieren in so genannte Kontexteffekte, die die Messung mitunter negativ beeinflussen können. Absichtliches Verfälschen („impression management“, „faking good“, sozial erwünschtes Antwortverhalten) in Persönlichkeitsfragebogen und Effekte unterschiedlicher Testreihenfolgen können beide als derartige Kontexteffekte bezeichnet werden. Die vorliegende Arbeit setzt sich aus drei wissenschaftlichen Artikeln zusammen, die den Einfluss unterschiedlicher Strategien (unterschiedliche Antwortformate, Instruktionen, limitierte Bearbeitungszeit und Fragebogenlänge) auf sozial erwünschtes Antwortverhalten, sowie den Effekt unterschiedlicher Testreihenfolgen auf die Testleistung in Objektiven Persönlichkeitstests und kognitiven Leistungstests an Bewerbern untersuchen. Die Ergebnisse der Studien werden ergänzt durch zwei weitere wissenschaftliche Artikel, die wiederum die Effekte unterschiedlicher Antwortformate auf sozial erwünschtes Antwortverhalten und die Effekte unterschiedlicher Testreihenfolgen auf die Testleistung untersuchen, jedoch an anderen Stichproben und mit teilweise anderen Verfahren. Die Ergebnisse unterstützen die Hypothese, dass Antwortskalen mit einer höheren Anzahl von Antwortalternativen (Analogskalen oder 6-kategorielle Rating-Skalen), Instruktionen mit der Warnung, dass Verfälscher entlarvt werden können, und Items, die am Ende eines Fragebogens platziert sind, zu weniger sozial erwünschtem Antwortverhalten führen, als Antwortskalen mit nur zwei Antwortalternativen (dichotome Antwortformate oder 2-kategorielle Rating-Skalen), Instruktionen ohne Warnung und Items, die am Beginn eines Fragebogens platziert sind. Diese Strategien scheinen für Testteilnehmer die Adaption ihrer Antworten an ein Faking-Schema zu erschweren (Holden & Hibbs, 1995; Holden, Kroner, Fekken, & Popham, 1992) oder deren Motivation zu verfälschen zu beeinflussen (Rothstein & Goffin, 2006). Eine limitiere Bearbeitungszeit führte ebenfalls zu einer Reduktion sozial erwünschten Antwortverhaltens, allerdings nur in Kombination mit einer Analogskala oder einer Warninstruktion. Die Effekte der einzelnen Strategien waren entweder innerhalb der einzelnen Studien oder über mehrere Studien hinweg nicht konsistent, was zu der Annahme führt, dass es eine Interaktion zwischen personenspezifischen Variablen und Skaleninhalten oder Itemformulierungen gibt. Weiters gibt es einen Einfluss unterschiedlicher Testreihenfolgen auf die Testleistung in Objektiven Persönlichkeitstests, nicht jedoch auf jene in kognitiven Leistungstests. Testteilnehmer, die zuerst an kognitiven Leistungstests und danach an Objektiven Persönlichkeitstest arbeiteten, zeigten eine höhere Entscheidungsfreude und eine geringere Frustrationstoleranz in den Objektiven Persönlichkeitstests, als Testteilnehmer, die zuerst an Objektiven Persönlichkeitstest

arbeiteten. Wiederum waren die Ergebnisse über mehrere Studien hinweg nicht konsistent. Insgesamt unterstützen die Ergebnisse die Theorie, dass individuelle Unterschiede das Verfälschen beeinflussen (Mueller-Hanson, Heggestad, & Thornton, 2006; Viswesvaran, & Ones, 1999) und die Theorie, dass Kontexteffekte von Unterschieden in der Motivation und in kognitiven Fähigkeiten von Testteilnehmern abhängen (Schwarz, Hippler, & Noelle-Neumann, 1992).

1. General Introduction

When we measure traits (characteristics, beliefs, abilities) by using psychological measures like questionnaires or tests, we should always be aware of the fact that responses are not always determined by mapping the trait directly on a response scale. As the relevant traits are predominantly not directly observable, we measure behaviours and reactions (provoked by psychological measures) that are assumed to be manifestations of these latent traits (Kubinger, 2009a). But in addition to these traits there can be other variables that influence or moderate the observable reactions by affecting the response process in different ways. Such moderator variables can be the presented content, the quality of test items, person-related variables (e.g. ability, motivation), or situation-related variables. They may result in numerous so called context effects, making measures more or less accurate or reliable. Intentional response distortion (in personality questionnaires or interviews) and order effects (with respect to all kinds of psychological measures) can be described as context effects because they result in altered response behaviour of test-takers that obscures the latent trait which was actually intended to be measured. Although the motivation behind understanding the causes of context effects is the desire to control their influence or to avoid their occurrence, context effects can also provide information about the processes involved in the generation of behavioural responses in experiments and surveys, and thus about human though processes (Bodenhausen, 1992). Therefore, the investigation of context effects can provide an important contribution to personality and social research. However, context effects can also harm the process of psychological assessment, particularly when it comes to situations where test results have consequences, like in personnel (or student) selection, where test scores of applicants are compared in order to identify the most qualified candidates.

The current thesis investigates both intentional response distortions (faking good, impression management) in personality questionnaires and order effects with respect to different orders of tests within a test battery. Because the majority of previous studies with regard to impression management and order effects have primarily used non-applicant volunteer samples, the studies of the present thesis use real job-applicants in personnel selection. The aim is to investigate different aspects of context effects in order to provide a contribution to personality research and social theories, as well as to models of response behaviour. Another aim is to investigate possibilities that could improve or optimise psychological assessment in personnel selection. *Paper 1* (as well as an additional study, presented below) investigates the effects of limited response time, response format, and warning instruction on intentional response distortion (impression management) in personality questionnaires. *Paper 2* addresses possible influences of questionnaire length on impression management. *Paper 3* is concerned with the effects of different test orders on test performance in cognitive ability tests and objective personality tests. Additionally

to these 3 papers, which form the current thesis, two additional papers are presented which expand their findings. *Paper 4* again investigates the influence of different response formats on intentional response distortion in personality questionnaires with respect to different kinds of rating scales. *Paper 5* attempts to discover whether the findings of Paper 3 can be replicated. All five studies comprise experiments that were conducted in selection situations to investigate the response behaviour of real applicants.

1.1. Faking

1.1.1. Faking good, faking bad: a definition of the common terminology

The term "faking" in psychological research is widely used to describe *intentional response distortion* in personality scales through the conscious, intentional description of oneself in a way that does normally not apply (cf. Franke, 2002). With respect to personnel selection, where it is assumed that applicants distort their responses to describe themselves in a positive way, we refer to the terms *faking good*, *impression management*, or *social desirability*. Contrarily, *faking bad* or *simulation*, where responses are distorted to present oneself in a negative way or to present an unrealistically negative impression, is rather assumed to occur in clinical settings or in forensic psychology (Franke, 2002; Kubinger, 2009a). The aim of both faking good and faking bad is to obtain a benefit or to avoid negative consequences. But the phenomenon of faking might not always be ascribed to intentional response distortion. It can also describe an unconsciously incorrect self-presentation due to a lack of self awareness or self perception, unconscious personality characteristics (cf. Franke, 2002), unclear instructions, or unclearly phrased test items. Moreover, faking good or socially desirable responding could be the result of a socialisation process where people are trained to present themselves in appropriate ways (Hogan, Barrett, & Hogan, 2007). According to this theory of impression management, faking represents socialized behaviour and occurs due to the fact that people might understand social norms better than their real disposition, measured by personality measures that mostly sample socialized adult behaviour. Therefore, *faking good* or *socially desirable responding* can be divided into *self-deceptive enhancement (SDE)* or self-deceptive tendencies, where test-takers believe in their positive self-reports, and *impression management (IM)* or self-favouring tendencies, where test-takers consciously distort their responses (Li & Bagger, 2006; Pauls & Crost, 2004; Pauls & Stemmler, 2003). Both tendencies can further be divided into an *egoistic bias*, where respondents see themselves as exceptionally talented or socially prominent, and a *moralistic bias*, where respondents see themselves as exceptionally good members of society (Paulhus & John, 1998; Pauls & Stemmler, 2003). While the egoistic bias (or Alpha; with the motive "need

for power") is assumed to be more related to self-deceptive enhancement and dimensions like "dominance", "extraversion", "openness", "neuroticism" or "emotional stability", "ambition", and "intellect", the moralistic bias (or Gamma; with the motive "need for approval") is assumed to be more related to impression management and dimensions like "dutifulness", "nurturance", "conscientiousness", and "agreeableness". According to research on impression management in interviews, impression management can be divided into different tactics used by applicants (Levashina & Campion, 2006): self-promotion tactics (where applicants intend to show that they possess desirable qualities for the job), ingratiation tactics (where applicants intend to evoke interpersonal liking and attraction), and defensive tactics (where applicants intend to protect or repair their image).

1.1.2. Faking good on personality questionnaires in personnel selection

Intentional response distortion (faking good, impression management) in personality questionnaires in a socially desirable or job-related desirable way is an interesting phenomenon in terms of how response behaviour is affected by different contexts resulting from situation-related (e.g. personnel selection), person-related (motivation and ability to fake, personality traits), and measure-related variables (e.g. content, presentation mode). Why personality questionnaires seem to be particularly vulnerable to response distortions, why organisations nevertheless show a growing interest in including personality questionnaires in their selection processes, what problems accompany this interest, and why research shows such controversy with regard to these problems, are questions discussed in the introductions of *Paper 1*, *Paper 2*, and *Paper 4*.

That applicants do fake is a well-documented phenomenon (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Griffin, Hesketh, & Grayson, 2004; Kanning & Holling, 2001; Karner, 2002; Kury, 2002; Robie, Brown, & Beaty, 2007; Thumin & Barclay, 1993). Considerable research has shown that even voluntary participants are able to intentionally fake good when instructed to empathize with a selection candidate (Krahé & Hermann, 2003; Kubinger, 2002; McFarland & Ryan, 2000; Schmit, Ryan, Stierwalt, & Powel, 1995; Winkelspecht, Lewis, & Thomas, 2006; Zickar & Robie, 1999) or to conform to a given job profile (Hoeth, Büttel, & Feyerabend, 1967; Lammers & Frankenfeld, 1999). But it was revealed that faking bad instructions lead to greater distortions than faking good instructions, and that applicants do not fake as much as volunteers under faking instructions (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Stumpf & Steinhart, 1981; Viswesvaran & Ones, 1999), though they obviously fake more than incumbents (Rosse, Stecher, Miller, & Levin, 1998). It was also shown that applicants do significantly elevate their scores when applying for a job in contrast to non-selection situations (Griffith, Chmielowski, & Yoshita, 2007). The different findings due to the use of applicant and non-applicant samples could be

explained by findings which reveal that test-takers show different faking styles or response patterns (Ellingson, Smith, & Sackett, 2001; Zickar, Gibby, & Robie, 2004). Instructing test-takers to fake generally leads to a highly socially desirable answer pattern across all items (and universally inflated mean scores) instead of affecting items differently depending on their content; this might not be an adequate approach to investigate effects of faking (Ellingson, Smith, & Sackett, 2001). Hence, findings of studies that used faking instructions on volunteers, instead of investigating applicants, might not be generalised to real-world selection settings. There are also concerns in using applicant-incumbent-comparisons to study faking, since no differences in personality scales between these two groups (Robie, Zickar, & Schmit, 2001) or an overlap of different response styles across these groups were found, with some applicants appearing to respond honestly and some incumbents appearing to fake their responses (Zickar, Gibby, & Robie, 2004). It can also be assumed that incumbents are less motivated to avoid careless responses in comparison to applicants (Dilchert, Ones, Viswesvaran, & Deller, 2006). Furthermore, it was shown that faking effects varied substantially across outcomes and selection situations; this leads to the suggestion that the extent to which faking might be a problem depends on test-takers' intentions and circumstances (Converse, Peterson, & Griffith, 2009). Thus, fakability is assumed to be a variable of individual differences (Mueller-Hanson, Heggestad, & Thornton, 2006; Viswesvaran & Ones, 1999). The introductions of *Paper 2* and *Paper 4* describe models which try to explain faking behaviour by showing how different moderating variables are linked together and interact with one another.

There are findings showing that faking in a personality questionnaire may be accompanied by the perception of its fakability (Steinmayr & Kersting, 2008). Therefore, it is possible that the perceived fakability of a measure leads to a low acceptance, and the low acceptance in turn leads to intentional response distortion. According to this theory, an increase of acceptance could lead to less intentional response distortion. For example, some authors argue that personality questionnaires in situations like personnel selection should be conceptualised as workplace simulations (Blickle, Momm, Schneider, Gansen, & Kramer, 2009). This assumes that, due to certain goal of self-presentation (work-self), not only responses to personality scales but also daily interactions at work are a "function of the interaction between the strength of the motives to get ahead and along and the degree of individual social skills". This assumption is further accompanied by findings that the criterion-related validities of work-specific contextualized personality scales were higher among incumbents instructed to respond as if applying for a very attractive job than among those instructed to respond honestly. Maybe the conceptualisation of personality questionnaires as workplace simulations is a way to enhance its acceptance.

1.1.3. Strategies to deal with faking

Strategies were revealed to deal with faking good, such as the identification of response distortion (with the use of response-time latencies or social desirability scales), the discouragement of test-takers from faking (with warning instructions that faking can be identified or will have negative consequences), or efforts to make personality questionnaires less fakable (through adjustment of the response format, the method of administration, or the position of items). *Paper 1*, *Paper 2* and *Paper 4* try to provide a contribution to such strategies and to models of impression management by investigating different administration methods. Their introductions also give an overview of the most interesting strategies with respect to the current thesis. Because strategies for dealing with impression management are of interest in the current thesis, a more detailed description of these strategies is provided in the following[2].

The aim of *identifying response distortion* is either to exclude applicants from the further selection process, or to control social desirability by correcting personality scores, a method that has no empirical evidence justifying its use (Goffin, & Christiansen, 2003) and that does not improve validity (Morgeson, Campion, Diboye, Hollenbeck, Murphy, & Schmitt, 2007). Actually, it was shown that the correction of personality scores was not able to reproduce the initial rank-order under honest answer conditions (Herzberg, 2004). Appropriate strategies are the use of response time latencies (Esser & Schneider, 1998; Holden & Hibbs, 1995; Holden, Kroner, Fekken & Popham, 1992; Hsu, Santelli & Hsu, 1989; Kuntz, 1974; Robie et al., 2000; Schneider & Hübner, 1980) and the use of specific scales to assess patterns of response distortion like social desirability, impression management, faking good, defensiveness, or self-deception (Crowne & Marlowe, 1960; Edwards, 1957; Hoeth, Büttel & Feyerabend, 1967; Paulhus, 1991; Schneider-Düker & Schneider, 1977). Both strategies are debateable as they do not necessarily identify intentional response distortions and participants who respond honestly could be eliminated as well (Hülsheger, Spinath, Küppers, & Etzel, 2004). Besides, social desirability scales (impression management as well as self-deceptive enhancement scales) are also fakable (Kury, 2002; Pauls & Crost, 2004; Pauls & Crost, 2005), which throws their usefulness to detect faking into question. It was also found that the score of a social desirability scale was not associated with lower agreement between self ratings and informant ratings (roommates, parents), but that it provided substantive trait information and should therefore not be used to determine the validity of measures or the veracity of self ratings (Kurtz, Tarquini, & Iobst, 2008). Moreover, not all social desirability scales seem to measure impression management tactics, but rather personality-related tendencies to view

[2] Some of the corresponding text passages are taken from the introductions of the *Papers 1, 2,* and *4*, but supplemented with additional or more detailed information.

oneself positively, which suggests the need for validity studies of such scales (Reid-Seiser & Fritzsche, 2001).

To *discourage test-takers from faking*, warning instructions (that faking can be identified or will have negative consequences) were used but were not altogether effective. Either significant effects on faking were limited to specific types of warnings (Dwight & Donovan, 2003), no effects of warning instructions were revealed at all (Kury, 2002), or warnings affected only particular kinds of measures (Ramsay, Schmitt, Oswald, Kim, & Gillespie, 2006). A warning about the presence of a socially desirable response subscale was shown to be effective (Honkaniemi & Feldt, 2008), while repeated visual presentations of different combinations of warning instruction and self-focused attention (presentation of the picture of an observing eye combined with a warning instruction) in a computer-based questionnaire were ineffective (Hülsheger, Spinath, Küppers, & Etzel, 2004). It was shown that a warning of response verification was associated with slower item response latencies, as it might have increased the complexity of response decisions (Vasilopoulos, Cucina, & McElreath, 2005). According to findings that test-takers' perceptions of the situation were strongly related to intentions to fake, it was hypothesised that the effects of warning instructions might be strengthened by altering perceptions about the importance of faking (e.g. "high scores are not necessarily desirable"), the efficacy of faking (e.g. "good detection methods are in place and fakers will be caught"), and subjective norms about faking (e.g. "faking is unacceptable"; Mueller-Hanson, Heggestad, & Thornton III, 2006). However, findings of lower mean scores in experimental conditions using a warning instruction may not necessarily reflect improved validity due to reduced intentional response distortion, but conservative responding instead (Converse et al., 2008; Dwight & Donovan, 2003). Another kind of instruction is to ask test-takers to answer accountably (which means to prove answers through behaviour in the future). There are findings showing that such an accountability instruction produced higher scores in the scales Conscientiousness and Emotional Stability in a big five measure, and it was hypothesised that higher validities of measures might be achieved with such an instruction (ter Laak, Leuven, & Brugman, 2000), a theory which needs to be supported by further research.

Strategies to make personality questionnaires less fakable focussed on aspects of how questionnaires are administered by adjusting the method of administration, the item positioning, or the response format.

Comparing computer-based questionnaires with paper-pencil questionnaires, or verbal with non-verbal questionnaires, revealed no differences (Amelang, Schäfer & Yousfi, 2002; Menghin & Kubinger, 1996; Richman, Kiesler, Weisband, & Drasgow, 1999).

Certain questionnaire scales ("neuroticism" and "conscientiousness") showed themselves to be less susceptible to socially desirable responding when items were randomly placed instead of grouped together (McFarland, Ryan, & Ellis, 2002). With respect to the effects of item position, it was shown that test-takers were more likely to

fake their answers at the beginning rather than at the end of a questionnaire, suggesting that they might have forgotten the faking instructions over time (Seiwald, 2002).

Covert item content or particular dimensions that have some positive as well as some negative sides, like "extraversion" (in contrast to dimensions that have almost negatively associations, like "neuroticism"), are assumed to be more difficult to fake (Furnham, 1986). Reducing specific contents of items in personnel selection which are particularly relevant to a job (like reducing the college relevance of items within a college selection process) are intended to make the items less fakable (Ramsay, Schmitt, Oswald, Kim, & Gillespie, 2006).

In contrast to items with a single-stimulus response format (or rating scale), items with a forced-choice response format are assumed to minimize faking tendencies, as they make it more difficult to respond desirably (Jackson, Wroblewski, & Ashton, 2000). These types of response formats may influence the perceived opportunity to fake, which in turn moderates the actual faking behaviour (Goffin & Boyd, 2009; Rothstein & Goffin, 2006; Snell, Sydell, & Lueke, 1999). *Single-stimulus response formats* are normative formats where the degree of agreement has to be marked on a rating scale (e.g. 1 = very inaccurate to 5 = very accurate); they allow interindividual comparisons (Heggestad, Morrison, Reeve, & McCloy, 2006). *Forced-choice formats* present two or more statements which appear equally attractive in order to assess different traits (Rothstein & Goffin, 2006), and are either ipsative or partially ipsative response formats (Heggestad, Morrison, Reeve, & McCloy, 2006). *Ipsative response formats*, where response alternatives have to be rated in relation to one another by marking one statement that is most and one statement that is least like another, only allow intraindividual comparisons. *Partially ipsative response formats* allow intraindividual as well as interindividual comparisons, as they provide characteristics of both ipsative and normative formats.

When test-takers were instructed to respond like applicants (fake good conditions), less faking was revealed for binary and quartet forced-choice formats, as well as ipsative and partially ipsative forced-choice formats, while single-stimulus response formats were more vulnerable to response distortions (Jackson, Wroblewski, & Ashton, 2000; Martin, Bown, & Hunt, 2002). Nevertheless, it has been shown that test-takers are able to distort their responses using a forced-choice format (Lammers & Frankenfeld, 1999). The forced-choice format was additionally shown to be a better predictor of personality and job-related abilities in fake good conditions than the single-stimulus format (Jackson, Wroblewski, & Ashton, 2000; Wright & Miederhoff, 1999). It was also revealed that both types of response formats were susceptible to response distortions, that subjects with higher cognitive ability were able to distort their answers more by using the forced-choice format than subjects with lower cognitive ability, and that a forced-choice response format was not better at retaining the rank ordering of individuals in comparison to a single-stimulus response format (Christiansen, Burns, & Montgomery, 2005; Heggestad, Morrison, Reeve, & McCloy, 2006). However, items with the forced-choice format showed higher construct

validity under fake good conditions (Christiansen, Burns, & Montgomery, 2005). With respect to criterion-related validity, no difference between these two response formats was revealed (Converse et al., 2008). As the samples in these studies consisted of volunteers (mostly students), the results still need to be demonstrated in real selection situations. In fact, one study compared real applicants with volunteers while applying different item formats (problem solving items, rote knowledge items, forced-choice items, self-description with rating scales, situational judgement items) and showed that participants could only distort items with a rating scale to their own advantage (Kanning & Kuhne, 2006)

Some research suggests that using analogue scales (in which participants mark the extent of their agreement along a continuous line between two alternatives) as a response format may be less prone to faking than a dichotomous (participants have to choose one of two alternatives) response format (Kubinger, 2002; Seiwald, 2002). It has also been suggested that a dichotomous response format provokes a kind of reactance resulting in untypical or arbitrary responses that do not describe the subject's true character (Karner, 2002).

1.2. Context effects

An overview of different context effects like carry-over and backfire effects or consistency and contrast effects, sequence or position effects, initial frame of reference effects, fatigue and learning effects, priming effects, clarifying and redefinition effects, logical connection effects, and focus effects is given by Smith (1992). Context effects are mainly investigated with respect to item and task order in questionnaires and achievement or cognitive ability tests. Order effects refer to the phenomenon that different orders of questions (or tasks) or response alternatives may influence test-takers' responses in a systematic fashion (Strack, 1992). The extent of this influence depends on the degree to which the content of the question, task, or response alternative determines the response. Prior items can determine how respondents interpret subsequent questions, thereby influencing the appropriate answers (Tourangeau, 1992). Of course, not only prior items themselves, but also how one responded to prior items may influence responses to later items (Smith, 1992). The interaction between prior responses and item order is called conditional order effect. However, there is hardly any information about the effects of different test orders within test batteries. Therefore, *Paper 3* and *Paper 5* investigated the influence of different test orders on test performance. An overview of the few findings about the influence of different test orders (most studies which deal with this topic are unpublished) is given in the introductions of *Paper 3* and *Paper 5*. Particularly carry-over effects, priming effects, and learning effects, which involve the transfer of prior content, meaning, or behaviour and

influence subsequent reactions, not to mention fatigue effects, might take place when using different test orders.

Carry-over effects (also termed *consistency effects* or *assimilation effects*) occur if prior items provide an interpretative framework for subsequent items during comprehension (Tourangeau, 1992). Later items are assumed to deal with the same topic or are drawn from the same category as previous items, even if there is no relation between the item contents. Such effects occur if topics are unfamiliar or stimuli are ambiguous. In contrast, if previous items do not provide an interpretative framework for a general item, intended as an overall summary of the more specific previous items, *backfire effects* (also termed *inconsistency effects* or *contrast effects*) may occur. This means that respondents think that the general item does not include material which was already covered by previous items (even if the test authors intended to include this material), maybe because of a desire to avoid redundancy. Backfire effects also appear if comparisons that lead to contrary judgements are made salient. For example, respondents might rate their present lives as less happy if they have recently recalled positive events from their past, or vice versa. Such backfire effects are termed retrieval-based backfire effects. There are also retrieval-based carry-over effects, where responding to prior items leaves material that is relevant to later items accessible to retrieval.

Priming effects occur if accessible categories (fuzzy sets of features organised around a prototype) or schemas (cognitive structures that represent knowledge about people, events, roles, the self, and the general processing of information) in memory are activated by features of a stimulus domain, directing attention and influencing how information is processed (Hogg & Vaughan, 2008; pp. 62-63). If a category is primed, stimuli are encoded by interpreting them in a category-consistent manner, but only if people do not know that they are primed or if they do not detect the cue (category). Otherwise, stimuli are interpreted in a category-incongruent manner. With regard to order effects, for example, the influence of previous questions on the responses to later questions can be understood as a priming event with an activation and information function (Strack, 1992). The activation function increases the accessibility of the activated information (resulting in an assimilation effect in the judgement) and does not require the respondent to be aware of the priming episode. The information function, in contrast, requires the respondent to be aware of the priming episode and to perceive an episodic relationship between the two questions (or that they share the same conversational context), resulting in assimilation or contrast effects. If the episodic relation between the questions is not perceived, but the respondent is aware of the priming episode, no assimilation or contrast effect might occur.

Learning effects are revealed when item reliabilities increase towards the end of a personality questionnaire (Hamilton, & Shuminsky, 1990; Hartig, Hölzel, & Moosbrugger, 2007; Knowles, 1988; Knowles & Byers, 1996; Knowles et al., 1992), or when higher test scores or lower item difficulties are observed towards the end of a cognitive ability test (Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007; Kubinger, 2009b). It has also

been shown that respondents do not only form a latent representation of psychometric instruments, but they do so very quickly (Ostrom, Betz, & Skowronski, 1992). This in turn guides their responses to the remaining items in the test. *Fatigue effects* due to later item positions result in more blanks towards the end of a questionnaire (Kraut, Wolfson, & Rothenberg, 1975), or they lead to slower reactions or increased errors in cognitive ability tests (Földényi, Tagwerker-Neuenschwander, Giovanoli, Schallberger, & Steinhausen, 1999). Item orders may influence a four-stage process (interpretation, retrieval of information, rendering a judgement, selection of a response) underlying response behaviour (Tourangeau, & Rasinski, 1988), which may not be limited to item order but may also extend to test order.

Order effects might interact with one another (Smith, 1992) and depend on a complex interaction of serial position, presentation mode, item attributes (e.g. plausibility, complexity, and extremity of the wording), and respondents' ability and motivation (Schwarz, Hippler, & Noelle-Neumann, 1992). Most of the studies focusing on order effects have not considered these variables and hence have not provided a satisfying explanation of order effects. *Paper 5* tried to take such variables into account by regarding differences in subjects' motivation and ability.

2. Contribution of the current doctoral thesis

2.1. Methodical approach

The methodical approach is an experimental one. Because studies which use faking instructions on volunteers might not be an adequate approach to investigate effects of intentional response distortion (Ellingson, Smith, & Sackett, 2001), all studies presented in the current thesis are based on experiments that were conducted in personnel selection or comparable situations (cf. *Paper 3* and *Paper 5*) to investigate the response behaviour of applicants. Specific variables (response format, time limit, warning instruction, questionnaire length, questionnaire content, test order) were varied and combined to investigate their effects on response behaviour and test performance. Applicants and test-takers were randomly assigned to the different experimental groups. Multivariate analysis of variance and Welch tests were conducted to compare the means of the experimental groups. In order to calculate the sample sizes needed to fulfil a priori precision requirements (type-I, type-II-risk, and relevant effect size) the program CADEMO (http://www.biomath.de) was used (the sample sizes for multivariate analysis of variance were calculated according to an analysis of variance design). All methods, analyses, and results are described in detail in the single papers (see 5.1., 5.2., 5.3., as well as Appendix 2). Summaries and additional literature are presented in the following.

2.2. The effect of speededness (time limit), response format, and warning instruction on impression management in personality questionnaires

2.2.1. Paper 1 and Paper 4

Because efforts to identify response distortions are debateable, *Paper 1* (Khorramdel & Kubinger, 2006; published in *Psychology Science*, latterly: *Psychological Test and Assessment Modeling*; see 5.1.) and *Paper 4* (Khorramdel & Kubinger; submitted to the *Journal of Personality Assessment*; see Appendix 2) concentrated on approaches to discourage test-takers from faking and on approaches to make personality questionnaires less fakable. According to Rothstein and Goffin (2006), even a small degree of ability to fake combined with the motivation to do so, as well as a small degree of motivation to fake combined with the ability to fake could already lead to response distortion. Therefore, an approach to reduce faking might be more successful if both ability and motivation to fake are considered. Such an approach was made in *Paper 1* by influencing subjects' ability to fake with an analogue scale as response format (in contrast to a dichotomous response format) and a set time limit (per questionnaire page) for response selection, as well as by

influencing subjects' motivation to fake with a warning instruction (in contrast to a standard instruction) in a completely crossed 2x2x2 design. Then, *Paper 4* again investigated the influence of different response formats (rating scales) on the ability to fake good. Both papers deal with different strategies to reduce impression management (or intentional response distortion).

In *Paper 1*, it was hypothesised that both a limited response time and an analogue scale as response format may decrease intentional response distortion by decreasing test-takers' ability to adjust their responses to a faking good schema (Holden & Hibbs, 1995; Holden, Kroner, Fekken, & Popham, 1992). That a warning instruction might increase the complexity of response decisions was shown by Vasilopoulos, Cucina, and McElreath (2005). The reason for this could be that responding more honestly might be more difficult and complex than matching responses with a stereotype faking schema. Furthermore, it was hypothesised that a repeated warning instruction (that faking can be detected) may lead to less response distortion by decreasing test-takers' motivation to fake. That a time limit can decrease the ability to fake is supported by the findings of Robie, Brown, and Beaty (2007) showing that faking responders take more time to complete their responses and make more corrections than honest responders. While *Paper 1* used forced choice items (MBTI) as well as single-stimulus (or normative) items (FKK) to investigate the difference between presenting them with an analogue or dichotomous scale, *Paper 4* used only single-stimulus items, as they are mostly used in personnel selection (because of their normative qualities, and because of fewer costs during their construction compared to forced-choice formats). The aim was to find out if the susceptibility of items with a single-stimulus response format (Jackson, Wroblewski, & Ashton, 2000; Martin, Bown, & Hunt, 2002) can be decreased by using a higher number of response alternatives (6-point rating scale) instead of using a minimum of two (2-point rating scale).

The multivariate analysis of variance in *Paper 1* revealed a significant main effect of the factor *response format* and significant interaction effect of the factors *response format* and *response time*.[3] Additional analyses (Welch tests)[4] of the scales "extroversion", "introversion", "feeling" and "thinking" (which were shown to be significant using Levene's test and therefore had to be excluded from the multivariate analysis of variance) revealed no significant effects ($p = .465$; $p = .535$; $p = .757$; $p = .882$). The multivariate analysis of variance in *Paper 4* revealed a significant main effect of the factor *response format* as well. See all results in more detail in *Paper 1* and *Paper 4*.

[3] The means and standard deviations for all scales are given by experimental condition completely in Table 4 in Appendix 1; the table in the published article is not complete because of an editorial mistake.
[4] As the results of the Welch tests are not provided in *Paper 1*, they are demonstrated in this text.

2.2.2. Additional Study

As 113 of the 208 participants of *Paper 1* were no real job-applicants (the remaining 95 participants are applicants from two personnel and management consulting companies), additional data of 27 job-applicants (of the same companies) were gathered by conducting the same experimental design (with the same measurements), and to repeat the analysis with the 95+27 = 122 applicants. The aim was to have a minimum 10 subjects in each experimental group, as a multivariate analysis of variance with $\alpha = .05$ and $\beta = .20$ is then able to detect a mean difference of $\delta \geq 2/3\ \sigma$ (the standard deviation of the test scores). The sample consisted of 74 women and 48 men between the age of 17 and 54 years; 24.8% of the applicants had a "lower" education (apprenticeship) and 75.2% had a "higher" education (general qualification for university entrance or university degree). They applied for different positions such as administrative, medical, and management positions. The applicants were distributed in the experimental groups of *Paper 1* as follows: Group 1, Group 2, Group 5, and Group 8 had 10 subjects each; Group 3 had 35 subjects; Group 4 had 11 subjects; Group 6 had 15 subjects; Group 7 had 21 subjects. Again, a multivariate analysis of variance was conducted to investigate the effects of the experimental conditions *response format* (analogue scale vs. dichotomous response format), *response time* (time limit vs. no time limit), and *instruction* (warning instruction vs. standard instruction), with additional attention to possible interaction effects of these conditions.

2.2.2.1. Results (Additional Study)

The means and standard deviations of all scales in each experimental condition are given in Table 1 in Appendix 1. After deleting the FKK scale "self concept of own competences", which was significant in the Levene's test ($p = .001$), Box's M-Test for testing the homogeneity of the variance-covariance matrix proved to be not significant ($p = .421$). That is, the resulting F-values of multivariate analysis of variance can be interpreted fairly. The multivariate analysis of variance showed a significant main effect of the factor *instruction* ($p < .037$; $F = 1.985$; $\eta^2 = .174$). The separate invariate analyses of the factor *instruction* for each single scale revealed significantly different means between the experimental groups in the two MBTI scales "feeling" ($p = .002$) and "thinking" ($p = .005$); the respective means are given in Table 2.

Table 2:
Means and standard deviations of the scales "feeling" and "thinking" at the different levels of the significant factor *instruction*

Scale	Instruction	Means	SD
Feeling	No Warning	7.511	2.989
	Warning	8.961	2.957
Thinking	No Warning	11.155	3.281
	Warning	9.532	3.246

Moreover, the multivariate analysis of variance showed a significant interaction effect between the factors *instruction* and *response time* ($p < .001$; $F = 3.121$; $\eta^2 = .248$). The separate invariate analyses of each single scale with regard to this interaction effect showed significantly different means among different experimental groups on the three MBTI scales "intuition" ($p = .004$), "sensing" ($p = .024$), and "introversion" ($p = .011$). See the respective means in Table 3.

Table 3:
Means and standard deviations of the scales "intuition", "sensing", and "introversion" at the different levels of the significant interactions of the factors *instruction* and *response time*

Scale	Instruction	Response Time	Means	SD
Intuition	No Warning	No Time Limit	7.650	2.739
		Time Limit	7.040	2.850
	Warning	No Time Limit	5.803	2.925
		Time Limit	8.904	3.793
Sensing	No Warning	No Time Limit	10.050	3.590
		Time Limit	10.840	3.891
	Warning	No Time Limit	12.767	4.134
		Time Limit	9.523	5.297
Introversion	No Warning	No Time Limit	6.150	3.674
		Time Limit	8.200	5.147
	Warning	No Time Limit	8.375	4.236
		Time Limit	5.904	3.534

To additionally investigate the effects of the factors *response format*, *response time*, and *instruction* on the FKK scale "self concept of own competences", which was removed from the multivariate analysis of variance, two-sample *t*-tests for unequal variances (Welch tests) were applied: while no significant effect occurred with regard to the factor *response*

format, significant effects were revealed with regard to the factor *instruction* ($p = .005$) and the factor *response time* ($p = .020$). The respective means are given in Table 4.

Table 4:
Means and standard deviations of the scale "self-concept of own competences" at the different levels of the factors *instruction* and *response time*

Scale	**Instruction / Response Time**	**Means**	**SD**
Self-concept of own competences	No Warning	1.266	1.483
	Warning	2.181	2.030
	No Time Limit	2.131	2.061
	Time Limit	1.369	1.481

Altogether, the multivariate analysis of variance revealed a main effect of the factor *instruction* and interaction effects between the factors *instruction* and *response time*. While the main effect of the factor *instruction* can be ascribed to significantly different group means in the two MBTI scales "feeling" and "thinking", the interaction effects of the factors *instruction* and *response time* concerns the three MBTI scales "intuition", "sensing", and "introversion". The following scores could generally be described as socially or occupationally desirable (see the descriptions of all MBTI scales in Table 1 of Paper 1): lower scores in the scales "feeling", "intuition", and "introversion", as well as higher scores in the scales "thinking" and "sensing".

According to the factor *instruction*, it was revealed that applicants who responded to the items of the scale "feeling" after receiving the warning instruction showed higher mean scores than those who received no warning instruction. Therefore, it can be assumed that the warning instruction led to less desirable responses compared to the standard instruction (no warning).

Through the interaction of the factors *instruction* and *response time*, it was shown that applicants who received a time limit as well as a warning instruction received higher scores in the scale "intuition" and lower scores in the scales "sensing" and "introversion" than applicants of the other experimental groups. With regard to the mean scores of the scales "intuition" and "sensing", it can be assumed that the combination of the time limit with the warning instruction led to less desirable responses than the other three combinations of the factors *instruction* and *response time* did (time limit with standard instruction, no time limit with warning instruction, no time limit with standard instruction) in the scales "intuition" and "sensing". In the scale "introversion", less desirable responses were revealed in the experimental group that received a warning instruction without a time limit.

The results of the Welch tests revealed that applicants who received a warning instruction or no time limit showed higher scores in the scale "self-concept of own

competences" than applicants who received a standard instruction or a time limit. As higher scores of this scale might be more desirable than lower scores, it can be assumed that applicants distorted their responses less when receiving a time limit than when no time limit was applied. However, the warning instruction showed no such effect. In contrast to the main effect of the multivariate analysis of variance, less desirable responses were revealed when no warning was applied.

2.2.3. Discussion and critical reflection (Paper 1, Paper 4, Additional Study)

Paper 1 showed that an analogue scale led to less socially (or occupationally) desirable responses than a dichotomous response format. Results of *Paper 4* revealed that a 6-point rating scale showed the same effects in contrast to a 2-point rating scale (less socially desirable responses) in most of the scales of another questionnaire. Comparing the findings of *Paper 1* with those of *Paper 4*, it is obvious that response distortions decrease with an increasing number of response alternatives. A higher number of response alternatives seem to make it more difficult to fit one's responses to an adopted faking good schema (Holden & Hibbs, 1995; Holden, Kroner, Fekken, & Popham, 1992). Interestingly, three scales of the questionnaire in *Paper 4* showed the opposite effect as a 2-point rating scale led to less socially desirable responses than the 6-point rating scale. In the discussion of *Paper 4*, it was theorised that there might be an interaction between the kind of response format and the item content as well as item wording. Furthermore, it was assumed that a 2-point rating scale could lead to an underestimation of the actual trait, thereby distorting the measurement.

In response to the hypothesis that a limited responding time might decrease response distortions in contrast to no time limit, *Paper 1* revealed that this is true only if the time limit is combined with an analogue scale. Considering the fact that in *Paper 1*, single-stimulus (or normative) items, as well as forced-choice items, were used, either presented with an analogue or dichotomous scale, it seems interesting that the effects of the factors *response format* and *response time* occurred only in questionnaire scales with single-stimulus items (FKK), while no effects occurred in scales with forced-choice items (MBTI). It might therefore be assumed that items with a single-stimulus response format (or rating scale) are less susceptible to faking when presented as an analogue scale than when presented as dichotomous items. But these two presentation types did not affect response behaviour when items were of a forced-choice-type format. *Paper 4* supports the findings of *Paper 1*, once again showing that rating scales (single-stimulus response formats) are less susceptible to impression management when more response alternatives are provided.

Contrary to the assumption that a warning instruction might reduce the motivation to fake and therefore decrease response distortions, results of *Paper 1* showed no effects.

However, the *Additional Study,* where the data of those subjects who were not real applicants (unemployed persons) were replaced with the data of additional real applicants, did. In this study, a warning instruction led to less socially desirable responses in two scales of the MBTI, while the reverse effect occurred in one scale of the FKK ("self-concept of own competencies"), where a standard instruction led to less socially desirable responses than a warning instruction. According to Mueller-Hanson, Heggestad, and Thornton III (2006), as well as Rothstein and Goffin (2006), the combination of a warning instruction with other strategies might strengthen the effect of the warning on reducing intentional response distortion by altering test-takers' perception of the efficacy of faking or their ability to fake. Though there were no interaction effects between the factor *instruction* and the factor *response format* or *response time* in *Paper 1*, these effects were found in the *Additional Study*. An interaction effect was revealed between the factors *instruction* and *response time*, showing that the combination of a warning instruction with a time limit led to less socially desirable responses in three scales (two scales of the MBTI and one scale of the FKK) than a warning instruction without time limit.

Beyond this combined effect, there was again no main effect of the factor *response time*. According to the group means of one scale (the MBTI scale "introversion"), a warning instruction with no time limit led to less socially desirable responses. Thus, the effects found in the *Additional Study* were not consistent. It is possible that different strategies to decrease impression management work differently, depending on the content of the questionnaire scales. In this respect, it might be important that the two scales ("introversion", "self-concept of own competencies") where either the time limit or the warning instruction did not work as hypothesized might have a higher affinity to somewhat clinical contents than the other scales, which were affected by the factors *instruction* and *response time*.

Apart from that, it is noticeable that the factor *response format* showed no effect on subjects' response behaviour in the *Additional Study* like it did in *Paper 1*, and that the effects of the warning instruction and the time limit occurred mainly in scales of the MBTI while the effects of *Paper 1* occurred only in scales of the FKK. This might be additional evidence for the assumption that such effects might depend on the content of the scales. Another explanation for the different findings might be, of course, the different samples; the subjects of *Paper 1*, who were not real applicants and who received a faking instruction, might have biased the results. The warning instruction (that intentional response distortion can be detected) in *Paper 1* might not have worked because these subjects did not have to expect negative consequences of their test results – in contrast to the subjects of the *Additional Study*, where the warning showed significant effects. While these differences between samples might explain the different findings of the effects of a warning instruction, they do not really explain the different findings with respect to the response format.

However, both the results of *Paper 1* and the results of the *Additional Study* have one similar effect in common: they showed that effects of time limit occurred only in combination with other strategies (either a particular response format or a warning instruction). According to the findings that respondents do form a latent representation of psychometric instruments very quickly (Ostrom, Betz & Skowronski, 1992), the time limit used in *Paper 1* and the *Additional Study* might not have been short enough to show an effect on response behaviour on its own. Further research could investigate other, more extreme time limits.

Conclusions

The current findings provide evidence for the models of impression management, which describe faking behaviour as an interaction of different variables (Goffin & Boyd, 2009; McFarland & Ryan, 2006; Mueller-Hanson, Heggestad, & Thornton, 2006). The effects of a 6-point rating scale or an analogue scale, as well as the effects of a warning instruction on intentional response distortion might be strengthened by other variables, or might enhance their effects. But this might not be true for all scales, as these effects seems to be bound to the scale or item content, as well as item wording. This assumption is supported by other studies which have found that items or scales are affected differently by intentional response distortion (or to a different extent), depending on their content (Ellingson, Smith, & Sackett, 2001) and on different faking or response styles (Zickar, Gibby, & Robie, 2004). The content of items or measurements as well as the kind of sample, should be given more consideration in further research on intentional response distortion, and items should be developed and used very carefully in personnel selection (with regard to their content and wording).

Limitations and implications for further research

The discussion of *Paper 1* mentions some limitations, which have to be considered with respect to the *Additional Study* as well. It should be noted that using an analogue scale or a 6-point rating scale is no guarantee for preventing intentional response distortions; they might only reduce the level of those distortions. As the effects of these response formats were either not consistent within and across the different studies, or occurred only with respect to particular scales of the questionnaires, they probably work in only a few settings or on specific personality questionnaires. The same limitations apply to the effects of the warning instruction and the time limit. Moreover, there is the opinion that lower mean scores due to warning instruction may not necessarily reflect improved validity due to reduced intentional response distortion, but conservative responding instead (Converse et al., 2008; Dwight & Donovan, 2003).

The assumption that these strategies work more or less depending on the item or scale content and on the type of sample is supported by different models of impression management (Goffin & Boyd, 2009; McFarland & Ryan, 2000; McFarland & Ryan, 2006; Mueller-Hanson, Heggestad, & Thornton, 2006; Snell, Sydell, & Lueke, 1999), but should be investigated further, for example, by varying the sample type and questionnaire content systematically. This might, however, be difficult with respect to the collection of data from applicants. As mentioned in the discussion of *Paper 1*, the unemployed participants, who were tested as part of a job application training programme, might have biased the results of *Paper 1*, even if they carried out a job application and had the possibility of experiencing how they would have performed in a personnel selection setting. This might also be an explanation for the different findings in the *Additional Study*, where the data of these participants were replaced with the data of applicants. The sample in *Paper 4* also constitutes a limitation to the results, as it is very unique (men who had all applied for the same training and who all came from the same institution); future research should investigate if the same effects can be found in other samples.

Future research should also identify item types or personality dimensions for which multidimensional scales (like an analogue scale or a six-point rating scale), a warning instruction, and a time limit work to reduce impression management. In this context, it would be interesting to study how different kinds of warning instructions (e.g. positive versus negative wording) and time limits (e.g. a time limit for each individual item versus a time limit for a set of items) affect response behaviour. The usefulness of limited response times is discussed in *Paper 1*, as a time limit might change the constructs being measured by a personality questionnaire.

The interpretation of scale means of the experimental groups might be a sufficient approach to studying response distortions if the aim is investigation of the effects of faking on different scales, given that item-level and scale-level analyses have identified effects on the same scales (Henry & Raju, 2006). In order to find out more about variables underlying or moderating response behaviour or different response styles, however, analysis on item-level would be interesting. After all, faking is a variable of individual differences (Mueller-Hanson, Heggestad, & Thornton, 2006; Viswesvaran, & Ones, 1999), and test-takers show different response or faking styles (Ellingson, Smith, & Sackett, 2001; Zickar, Gibby, & Robie, 2004). Future research should apply IRT-based analysis or Mixed Rasch Models, if a higher number of applicants is available; Latent Class Analysis would also be interesting.

Another point of critique could be seen in the fact that we did not use a control group. The most appropriate control group with respect to applicants would, of course, be incumbents who already have the job that applicants are applying for. But there are actually concerns about using applicants and incumbents to study faking, as no differences in personality scales between these two groups (Robie, Zickar, & Schmit, 2001) or an overlap of different response styles across these groups were found; some applicants

appeared to respond honestly and some incumbents appeared to fake their responses (Zickar, Gibby, & Robie, 2004).

2.3. Questionnaire length and impression management

2.3.1. Paper 2

While *Paper 1* and the *Additional Study* investigated the influence of a limited response time on response behaviour in questionnaires, *Paper 2* (Khorramdel, Kubinger, & Uitz; submitted to *International Journal of Selection and Assessment*; see 5.2.) investigated the influence of item position with respect to the questionnaire length. The aim is to provide a contribution to research on intentional response distortion, in particular regarding the influence of different contexts on impression management. Furthermore, the influence of the length of a questionnaire on different kinds of questionnaires with either work-related (BIP) or no specific work-related contents (NEO FFI, NEO PI-R) was the focus of interest. A questionnaire with a 6-point rating scale consisting of 516 items from different (well known) questionnaires was administered to 84 applicants from the Federal Armed Forces, who had applied for pilot training. The positions of the 516 items were varied to test if responses are affected by an overlong test length.

It was hypothesised that socially desirable response distortion would either increase or decrease towards the end of a very long questionnaire as learning or fatigue effects might occur (Hartig, Hölzel, & Moosbrugger, 2007; Knowles, 1988; Seiwald, 2002) making it more or less easy to adjust responses to a faking good schema (Holden & Hibbs, 1995; Holden, Kroner, Fekken, & Popham, 1992). It was assumed that increasing response distortion towards the end of the questionnaire (learning effects) would result in decreasing scale reliabilities as responses of the sample become more stereotypical, and that decreasing intentional response distortion towards the end of a questionnaire (fatigue effects) would result in increasing scale reliabilities as responses of the sample become more variable and less stereotypical. Furthermore, it was hypothesised that a questionnaire that measures the big five dimensions might be more vulnerable to response distortion than a questionnaire with work-related content, as socially desirable responding seems to affect particular dimensions like Neuroticism, Emotional Stability, Agreeableness, and Conscientiousness (McFarland & Ryan, 2000; Ones, Viswesvaran & Reiss, 1996; Rosse, Stecher, Miller, & Levin, 1998).

Results from the multivariate analysis of variance showed a significant main effect of the factor *item position*. The reliabilities of most of the scales showed higher reliabilities at the end of the questionnaire than when they were applied at the beginning of the questionnaire, except for two scales. See all results in more detail in *Paper 2*.

2.3.2. Discussion and critical reflection (Paper 2)

In line with literature (Seiwald, 2002), the findings of *Paper 2* showed that a fatigue effect might have occurred due to the extensive length of the administered questionnaire, decreasing the concentration or alertness towards the end of the questionnaire, thereby also decreasing the applicants' ability to adjust their responses in questionnaires to an adopted faking good schema. Although, only one scale ("conscientiousness") was shown to be affected by the extensive test length, almost all scales, except for two, show a common trend: their scale reliabilities tend to be higher at the end rather than at the beginning of the questionnaire, showing that socially desirable responses decreased towards the end of the questionnaire (fatigue effect). The results of the two scales ("agreeableness", "angry hostility"), which showed a different trend, were interpreted as a kind of frustration effect rather than as an expression of faking tendencies (or learning tendencies). Again, it seems as if not all scales or items are affected by intentional response distortion to the same extent, which is supported by other studies showing that items or scales are affected differently by intentional response distortion (or to a different extent) depending on their content (Ellingson, Smith, & Sackett, 2001) or different faking or response styles (Zickar, Gibby, & Robie, 2004).

Furthermore, (again in line with common literature) it was revealed that the big five dimension "conscientiousness" from the NEO FFI seems to be more vulnerable to intentional response distortions than other big five dimensions or the work-related scales of the BIP. This finding might, of course, also be an artefact, but with regard to the literature (McFarland & Ryan, 2000; Ones, Viswesvaran & Reiss, 1996; Rosse, Stecher, Miller, & Levin, 1998) it can rather be assumed that the items of the dimension "conscientiousness" might be more transparent with regard to the measured content than items from other big five dimensions, making it more easy to choose socially desirable responses at the beginning of a questionnaire when concentration and alertness are still given. However, this does not mean that the scales of the BIP are not affected by intentional response distortion.

Limitations and implications for further research

The discussion in *Paper 2* discloses certain limitations that future research should address. One limitation is, like in *Paper 4*, the sample, which is very unique (men, who had all applied for the same training and who all came from the same institution). Therefore, the findings might not apply to other groups, such as women, or other occupational groups. Future research should investigate if the same effects can be found in other samples.

As already discussed above (with respect to the findings from *Paper 1*, the *Additional Study*, and *Paper 4*; see 2.2.3.), we cannot maintain that all applicants distorted their responses or that all applicants actually distorted their responses to the same extent. Using

a control group of incumbents to find out if the applicants did fake might not be a satisfying approach because incumbents fake as well (Zickar, Gibby, & Robie, 2004). Again, item based analysis, such as an IRT analysis or Mixed Rasch Models, but also a Latent Class Analysis, would be interesting in further research (if a larger sample size is available) as more information might be revealed with regard to response behaviour or different response styles.

Furthermore, the discussion in *Paper 2* mentions the need to ascertain the hypothesis, that fatigue effects decrease the ability to fake by making it more difficult to fit responses to a faking schema, in further research by measuring item response latencies. Finally, the findings might have more relevance for research than for practice in personnel selection, as it is not clear if such fatigue effects alter the constructs measured by questionnaires by reducing the ability to give information about one's real characteristics.

2.4. The effect of test order on test performance (with special regard to objective personality tests)

2.4.1. Paper 3 and Paper 5

While other measures such as situational judgement tests (Peeters & Lievens, 2005), interviews (Levashina & Campion, 2006), and biodata (biographical information) measures (Kluger, Reilly, & Russell, 1991; Levashina, Morgeson, & Campion, 2009) are fakable as well, objective personality tests sensu R. B. Cattell (e.g. 1958) present a promising alternative to personality questionnaires with respect to their use in selection settings as they are less vulnerable to intentional response distortion (Baldinger, 2006; Hofmann & Kubinger, 2001; Ziegler, Schmidt-Atzert, Bühner, & Krumm, 2007). One reason might be that the computation of test scores is not transparent to test-takers. Objective personality tests are experiment-based assessments of behaviour, which assess a personality construct by observing the subject's behaviour when working on a performance or ability task, while the behaviour is observed and registered on a computer (Kubinger, 2009c). They could be described as individual computerised assessments, which oftentimes include computer simulations of job-related tasks. *Paper 3* (Khorramdel & Frebort; accepted for publication in *European Journal of Psychological Assessment*; see 5.3.) and *Paper 5* (Khorramdel & Frebort; submitted to *European Journal of Psychological Assessment*; see Appendix 2) investigated if objective personality tests are vulnerable to other context effects than impression management like different test orders. The use of different test orders within test batteries is a common practice for different reasons (see the introduction of *Paper 3*), but not well explored or proven to be without consequences for the test results. Therefore, *Paper 3* presents an experiment where the sequence of objective

personality tests and cognitive ability tests was varied within a computer based test battery. The sample consisted of 66 managers in an industrial corporation (an automotive supplier) in "higher management positions" (business managers, department chiefs, and team leaders), who underwent an analysis of their professional potential which resembles a real selection situation. In *Paper 5* the same experiment was performed by testing 64 incumbents of the same corporation, who were in "lower positions" (shift foremen and machinery adjusters) but who still had managerial responsibilities. In contrast to *Paper 5* a different matrices test was administered. The aim was to investigate the effects of a varied test order in different kind of samples with respect to possible differences in cognitive ability, achievement motivation, and resilience.

It was hypothesized that carry-over and priming effects, as well as fatigue and learning effects might occur. With regard to the model from Petty and Cacioppo (1986), as well as that from Schwarz, Hippler, and Noelle-Neumann (1992), it was further assumed that effects of test order might occur depending on differences in subjects' cognitive ability and motivation.

The multivariate analysis of variance in *Paper 3* showed a main effect of the factor *test order*, which could not be replicated in *Paper 5*, where no main effect of the factor *test order* occurred. See all results in more detail in *Paper 3* and *Paper 5*.

2.4.2. Discussion and critical reflection (Paper 3, Paper 5)

It was revealed that different test orders have a significant effect on test performance as subjects who worked on ability tests first and on objective personality tests second showed increased "decisiveness" and lower "frustration tolerance" scores in the objective personality test *Work Styles* in comparison to subjects who worked on objective personality tests first (showing fatigue effects on the one hand, and the vulnerability of particular kinds of resilience measured with tasks on the other hand). However, these effects only occurred in the sample from *Paper 3,* which exhibited higher cognitive abilities and status, as well as higher aspiration levels and lower resilience (in the objective personality tests) than the sample from *Paper 5*. Moreover, the effects of the test order only occurred in subtests with very simple tasks (such as the Work Styles), while more complex tasks (such as AMT, SPM, IST 2000 R) were not affected. The discussion in *Paper 3* provides a comparison of the findings with similar findings in literature. The discussion of *Paper 5* provides possible explanations for the different findings in *Paper 3* and *Paper 5*, such as differences between the two samples as well as differences in the two test batteries (in *Paper 3* the AMT matrices test was administered, while the SPM matrices test was used in *Paper 5*).

Limitations and implications for future research

In the discussion in *Paper 3* and *Paper 5* some limitations regarding the findings are discussed and implications for further research are provided. The findings of both studies in these papers might not apply to other groups apart from the investigated samples, or may not be generalised, as the samples were again unique ones (most participants were men and the participants were all managers in a particular industrial corporation). The interactions between sample, task type, and content of the test should be further explored. For example, it was revealed that different test orders affected only particular kinds of resilience measured by tasks, as corresponding scores in the *Work Styles* were affected but none of the scores in BAcO. The possible effects of the different matrices tests on the test scores from *Work Styles* should be investigated separately from other moderating variables (such as differences in motivation or cognitive ability). Furthermore, it would also be interesting to investigate possible order effects by varying the order of different objective personality tests; their order was held constant in the current experiment.

3. General Discussion and Prospects

Altogether, the different studies in the papers showed that response behaviour in personality questionnaires and reactions in objective personality tests are influenced by different contexts, such as different presentation modes (response format, instruction, response time, questionnaire length) and different test orders. Response scales such as analogue scales and rating scales with a higher number of response alternatives (6-point rating scale) showed less socially (or job) desirable responses in questionnaires than response formats with only two response alternatives (dichotomous response format, 2-point rating scale). Decreased response distortion was also revealed by using instructions which include the warning that fakers can be detected, in comparison to instructions without such a warning. A limitation of response time (per questionnaire page) showed similar effects but only when combined with an analogue scale or a warning instruction, which leads to the assumption that a time limit might strengthen other strategies, but has no effect on its own. Moreover, an extensive questionnaire length was shown to result in fatigue effects, reducing test-takers' ability to fake. The latter finding might have more relevance for research than for practice in personnel selection, as it is not clear if such fatigue effects alter the constructs measured with questionnaires by reducing the ability to give information about one's real characteristics as well. Similar limitations can be assumed when the response time is reduced in questionnaires. Further research should investigate if a time limit reduces the comprehension of items or disadvantages test-takers who are less resilient (working under pressure, coping with stress). Moreover, the findings were either not consistent within and across the different studies, or occurred only with respect to particular scales. Therefore, the different strategies might probably work in only a few settings or personality questionnaires because items or scales are affected differently by intentional response distortion (or to a different extent) depending on their content (Ellingson, Smith, & Sackett, 2001) or different faking styles (Zickar, Gibby, & Robie, 2004). Objective personality tests are less vulnerable to intentional response distortions than personality questionnaires (Baldinger, 2006; Hofmann & Kubinger, 2001; Ziegler, Schmidt-Atzert, Bühner, & Krumm, 2007), but seem to be affected by different test orders with regard to particular scores such as "decisiveness" and "frustration tolerance" (*Work Styles*). Again, the findings were not consistent across different studies.

There are, of course, certain limitations with regard to the different findings, which are discussed in the papers as well as in the text above. As the studies, which are presented in the different papers, were conducted in a personnel selection situation (*Paper 1, Paper 2, Paper 4, Additional Study)* or similar situations (*Paper 3, Paper 5*), the advantage, being that applicants and not volunteer test-takers were investigated, allows conclusions for the use of personality questionnaires and objective personality tests in selection situations.

Although, the context of a selection situation was given in all studies, the differences between the samples (and maybe the different selection situations) had a noteworthy influence on the findings as results were not consistent across the different studies. According to Zickar, Gibby, and Robie (2004), a variety of faking styles might have occurred in *Paper 1*, *Paper 2*, and *Paper 4*, influencing the current results. Therefore, evidence is provided for the theory that faking is a variable of individual differences (Mueller-Hanson, Heggestad, & Thornton, 2006; Viswesvaran, & Ones, 1999), which also depends on the kind of situation and circumstances (Converse, Peterson, & Griffith, 2009). Likewise, the findings of *Paper 3* and *Paper 5* provide evidence for the theory that the appearance and the extent of context effects depend on differences in subjects' motivation and cognitive ability (Schwarz, Hippler, & Noelle-Neumann, 1992). Moreover, the current findings allow the assumption that the contents of the measurements are affected differently by the experimental conditions because particular questionnaire scales and test scores appeared to be more vulnerable to intentional response distortion or test order effects. Hence, one important finding of the current thesis is that there are nameable interactions between sample (person related variables), measurement (test related variables, such as task type, item content, and wording) and situation related variables that should not be neglected. There are different strategies to deal with intentional response distortion which seem to be promising, but do not solve the problem of rank order changes in personnel selection due to intentional response distortion as they do not seem to work equal for all scales or all persons, so that no general conclusions can be drawn. In fact, it seems as if every strategy and combination between different strategies would need to be investigated in all interesting samples and situations for all interesting questionnaires. The same applies for the effects of different test orders. Future research should address the influence of different person related variables and the content of different personality measures.

The problem with faking might not be faking itself, but differences in the extent of faking between applicants who show different faking styles (Ellingson, Smith, & Sackett, 2001; Zickar, Gibby, & Robie, 2004); some might even respond honestly, depending on variables such as their ability and motivation to fake, or on their real personality characteristics (McFarland & Ryan, 2000; McFarland & Ryan, 2006; Mueller-Hanson, Heggestad, & Thornton, 2006; Snell, Sydell, & Lueke, 1999). Different faking or response styles, due to differences in the motivation to fake, mean that each applicant has a different starting point. This, in turn, means that the measurement is, of course, not fair. According to the belief that intentional response distortion is an expression of social competence that could predict job performance (Marcus, 2003a, 2003b; Morgeson, Campion, Diboye, Hollenbeck, Murphy, & Schmitt, 2007) and the idea that personality questionnaires should be conceptualised as work place simulations (Blickle, Momm, Schneider, Gansen, & Kramer, 2009), there might be a possibility for the use of personality questionnaires in personnel selection anyhow. Applicants should have the possibility to start from the same

point, which could be achieved by avoiding different motivations to respond to personality questionnaires. Personality questionnaires should be introduced as measures to assess the "knowledge of which behaviour is appropriate or desirable in work or society, or the "knowledge of which behaviour is successful in work". Therewith, rank order changes would only depend on the applicants' knowledge and ability, which might be variables that predict job performance. In addition, scales should be used which measure job-related traits and which are transparent with regard to their measured content as personality scales might provide substantial criterion validity in personnel selection if the measured traits are matched to the nature of the job (Goffin & Boyd, 2009). Of course, questionnaire scales would then not measure the "choice to perform" or "will do" aspect (Goffin & Boyd, 2009) but rather the "knowledge of performance" or "know how to do" aspect. They might also measure whether applicants have a realistic expectation of the job, which might also be a factor of success. Further studies should investigate possible benefits or problems of such a procedure in personnel (or student) selection and the effects on the validity of measures.

An interesting alternative to the use of questionnaires is still objective personality tests, with their advantage of measuring the "capacity" to perform. However, they have the disadvantage of only being able to assess particular personality characteristics. When using objective personality tests it needs to be taken into consideration that order effects may or may not occur, depending on the sample, the kind of objective personality tests or tasks, and other measurements which might be included in the same test battery. As there is hardly any knowledge of the effects of different test orders on the test scores in objective personality tests at present, it might be advisable to hold the test order of test batteries, which comprise objective personality tests or comparable computer simulations, constant for all participants. This might be particularly relevant in personnel selection, where the test scores of applicants are compared in order to identify those who are most qualified. On the other hand, if certain effects due to test order are intended, they might be controllable and therefore beneficial, if more is known about their occurrence.

4. References

Amelang, M., Schäfer, A., & Yousfi, S. (2002). Comparing verbal and non-verbal personality scales: Investigating the reliability and validity, the influence of social desirability, and the effects of fake good instructions. *Psychologische Beiträge* [latterly: *Psychological Test and Assessment Modeling*]*, 44,* 24-41.

Baldinger, D. (2006). *Der Einfluss sozial erwünschten Verhaltens auf das Ergebnis Objektiver Persönlichkeitstests [The influence of social desirable behaviour on the results of objective personality questionnaires]*. Unpublished diploma thesis, University of Vienna.

Birkeland, S.A., Manson, T.M., Kisamore, J.L., Brannick, M.T. & Smith, M.A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment, 14*(4), 317-335.

Blickle, G., Momm, T., Schneider, P. B., Gansen, D., & Kramer, J. (2009). Does acquisitive self-presentation in personality self-ratings enhance validity? Evidence from two experimental field studies. *International Journal of Selection and Assessment, 17*, 142-260.

Bodenhausen, G.V. (1992). Information-Processing Functions of Generic Knowledge Structures and Their Role in Context Effects in Social Judgment. In N. Schwarz & S. Sudman. *Context effects in Social and Psychological Research,* pp. 267-277. New York: Springer.

Cattell, R. B. (1958). What is "objective" in "objective" personality tests? *Journal of Counseling Psychology, 5*, 285-289.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item format for applicant personality assessment. *Human Performance, 18*, 267-307.

Converse, P. D., Oswald, F. L., Imus, A., Hedricks, C., Roy, R., & Butera, H. (2008). Comparing personality test formats and warnings: Effects on criterion-related validity and test-taker reactions. *International Journal of Selection and Assessment, 16*, 155-169.

Converse, P. D., Peterson, M. H., and Griffith, R. L. (2009). Faking on personality measures: Implications for selection involving multiple predictors. *International Journal of Selection and Assessment, 17*, 47-60.

Crowne, D.P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24,* 349-354.

Dilchert, S., Ones, D. S., Viswesvaran, C., & Deller, J. (2006). Response distortion in personality measurement: born to deceive, yet capable of providing valid self-assessments? *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*]*, 48*, 209-225.

Dwight, S. A. & Donovan, J. J. (2003). Do warnings not to fake actually reduce faking? *Human Performance, 16*, 1-23.

Edwards, A.L. (1957). *The social desirability variable in personality assessment and research.* New York, NY: Dryden Press.

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.

Esser, C., & Schneider, J.F. (1998). Differentielle Reaktionslatenzzeiten beim Bearbeiten von Persönlichkeitsfragebogen als möglicher Indikator für Verfälschungstendenzen [Differential response latencies as a possible indicator for detecting faking on personality test items]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 19,* 246-257.

Franke, G. H. (2002). Faking bad in personality inventories: Consequences for the clinical context. *Psychologische Beiträge* [latterly: *Psychological Test and Assessment Modeling*]*, 44*, 50-61.

Földényi, M., Tagwerker-Neuenschwander, F., Giovanoli, A., Schallberger, U., & Steinhausen, H.-C. (1999). Die Aufmerksamkeitsleistungen von 6-10-jährigen Kindern in der TAP [Attentional performance of 6-10-year-old children on the TAP]. *Zeitschrift für Neuropsychologie, 10,* 87-102.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences, 7*, 385-400.

Goffin, R. D., & Boyd, A. C. (2009). Faking and personality assessment in personnel selection: Advancing models of faking. *Canadian Psychology, 50*, 151-160.

Goffin, R.D., & Christiansen, N.D. (2003). Correcting Personality Tests for Faking: A Review of Popular Tests and an Initial Survey of Researchers. *International Journal of Selection and Assessment, 11,* 340-344.

Griffin, B., Hesketh, B., & Grayson, D. (2004). Applicants faking good: evidence of item bias in the NEO PI-R. *Personality and Individual Differences, 36*, 1545-1558.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behaviour. *Personnel Review, 36*, 341-355.

Hamilton, J. C., & Shuminsky, T. R. (1990). Self-awareness mediates the relationship between serial position and item reliability. *Journal of Personality and Social Psychology, 59, 1301-1307.*

Hartig, J., Hölzel, B., & Moosbrugger, H. (2007). A Confirmatory Analysis of Item Reliability Trends (CAIRT): Differentiating true score and error variance in the analysis of item context effects. *Multivariate Behavioral Research, 42,* 157-183.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*, 373-385.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*, 9-24.

Henry, M. S., & Raju, N. S. (2006). The effects of traited and situational impression management on a personality test: an empirical analysis. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*], *48*, 247-267.

Herzberg, P. Y. (2004). Lässt sich der Einfluss sozialer Erwünschtheit in einem Fragebogen zur Erfassung aggressiver Verhaltensweisen im Straßenverkehr korrigieren [Correction of social desirability distortion in a questionnaire of aggressive traffic behaviour]? *Zeitschrift für Differentielle und Diagnostische Psychologie, 25*, 19-29.

Hoeth, F., Büttel, R., & Feyerabend, H. (1967). Experimentelle Untersuchungen zur Validität von Persönlichkeitsfragebogen [Experimental investigation on the validity of personality questionnaires]. *Psychologische Rundschau, 18,* 169-184.

Hofmann, K., & Kubinger, K. D. (2001). Herkömmliche Persönlichkeitsfragebogen und Objektive Persönlichkeitstests im „Wettstreit“ um Unverfälschbarkeit [Personality questionnaires and objective personality tests in contest: More or less fakeable?]. *Report Psychologie, 26,* 298-304.

Hogan, J., Barrett, P. & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92*, 1270-1285.

Hogg, M. A., & Vaughan, G. M. (2008). *Social Psychology*. Fifth edition. Harlow, UK: Pearson.

Holden, R., & Hibbs, N. (1995). Increment validity of response latencies for detecting fakers on a personality test. *Journal of Research in personality, 29,* 362-372.

Holden, R. R., Kroner, D. G., Fekken, G. C., & Popham, S. M. (1992). A model of personality test item response dissimulation. *Journal of Personality and Social Psychology, 63,* 272-279.

Honkaniemi, L., & Feldt, T. (2008). Egoistic and moralistic bias in real-life inventory responses. *Personality and Individual Differences, 45,* 307-311.

Hough, L.M., Eaton, N.K., Dunnette, M.D., Kamp, J.D., & McCloy, R.A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581-595.

Hsu, L.M., Santelli, J., & Hsu, J.R. (1989). Faking detection validity and incremental validity of response latencies to MMPI Subtle and Obvious item. *Journal of Personality assessment, 53,* 278-295.

Hülsheger, U.R., Spinath, F.M., Küppers, A., & Etzel, S. (2004). Experimentelle Untersuchung zweier Methoden zur Reduzierung Sozialer Erwünschtheit in einem computergestützen eignungsdiagnostischen Testverfahren [Experimental study on two methods to reduce social desirability in a computer-based test for personnel selection and development]. *Zeitschrift für Personalpsychologie, 3,* 24-33.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced-choice offer a solution? Human Performance, 13, 371-388.

Kanning, U. P., & Holling, H. (2001). Struktur, Reliabilität und Validität des NEO-FFI in einer Personalauswahlsituation [Structure, reliability, and validity of the NEO-FFI in a personnel selection process]. *Zeitschrift für Differentielle und Diagnostische Psychologie. 22,* 239-247.

Kanning, U. P., & Kuhne, S. (2006). Social desirability in a multimodal personnel selection test battery. *European Journal of Work and Organizational Psychology, 15*, 241-261.

Karner, T, (2002). The volunteer effect of answering personality questionnaires. *Psychologische Beiträge* [latterly: *Psychological Test and Assessment Modeling*], *44,* 42-49.

Kluger, A. N., Reilly, R. R., & Russel, C. J. (1991). Faking biodata tests: Are option-keyed instruments more resistant? *Journal of Applied Psychology, 76*, 889-896.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*, 312-320.

Knowles, E. S., & Byers, B. (1996). Reliability shifts in measurement reactivity: Driven by content engagement or self-engagement? *Journal of Personality and Social Psychology, 70*, 1080-1090.

Knowles, E. S., Coker, M. C., Cook, D. A., Diercks, S. R., Irwin, M. E., Lundeen, E. J., Neville, J. W., & Sibicky, M. E. (1992). Order effects within personality measures. In N. Schwarz & S. Sudman. *Context effects in Social and Psychological Research,* (pp. 221-2236). New York: Springer.

Krahé, B., & Herrmann, J. (2003). Verfälschungstendenzen im NEO-FFI: Eine experimentelle Überprüfung [Faking on the NEO-FFI: An experimental investigation]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 24*, 105-117.

Kraut, A. I., Wolfson, A. D., & Rothenberg, A. (1975). Some effects of position on opinion survey items. *Journal of Applied Psychology, 60*, 774-776.

Kubinger, K. D. (2002). On faking personality inventories. *Psychologische Beiträge* [latterly: *Psychological Test and Assessment Modeling*]*, 44,* 10-16.

Kubinger, K. D. (2009a). *Psychologische Diagnostik: Theorie und Praxis psychologischen Diagnsotizierens* [*Psychological assessment: theory and practice of psychological diagnostics*]. Wien: Hogrefe.

Kubinger, K. D. (2009b). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement, 69*, 232-244.

Kubinger, K. D. (2009c). The technique of objective personality-tests sensu R. B. Cattell nowadays: The Viennese pool of computerized tests aimed at experiment-based assessment of behaviour. *Acta Psychologica Sinica*, *41*, 1024-1036.

Kuntz, D. (1974). Effects of faking instructions on the word-association test. *Psychological Reports, 35,* 1183-1192.

Kurtz, J. E., Tarquini, S. J., & Iobst, E. A. (2008). Socially desirable responding in personality assessment: Still more substance than style. *Personality and Individual Differences, 45,* 22-27.

Kury, H. (2002). Das Freiburger Persönlichkeitsinventar und sein Einsatz bei kriminologischen Fragestellungen. Das Problem der Verfälschungstendenzen [The Freiburger Personality Inventory and its use for criminological questions. The problem of faking tendencies]. In M. Myrtek (Hrsg.). *Die Person im biologischen und sozialen Kontext* [*The person within a biological and social context*] (S.249-270). Göttingen: Hogrefe.

Lammers, F., & Frankenfeld, V. (1999). Effekte gezielter Antwortstrategien bei einem Persönlichkeitsfragebogen mit „forced-choice"-Format [Effects of selective response strategies in a personality questionnaire with forced-choice format.] *Diagnostica, 45,* 65-68.

Levashina, J., & Campion, M. A. (2006). A model of faking likelihood in the employment interview. *International Journal of Selection and Assessment, 14,* 299-316.

Levashina, J., Morgeson, F. O., & Campion, M. A. (2009). They don't do it often, but they do it well: Exploring the relationship between applicant mental abilities and faking. *International Journal of Selection and Assessment, 17,* 271-281.

Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment, 14,* 131-141.

Marcus, B. (2003a). Das Wunder sozialer Erwünschtheit in der Personalauswahl. [The miracle of social desirability in personnel selection settings] *Zeitschrift für Personalpsychologie, 2,* 129-132.

Marcus, B. (2003b). Persönlichkeitstests in der Personalauswahl: Sind „sozial erwünschte" Antworten wirklich nicht wünschenswert? [Personality testing in personnel selection: Is „socially desirable" responding really undesirable?]. *Zeitschrift für Personalpsychologie, 2,* 138-148.

Martin, B. A., Bown, C.-C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences, 32,* 247-256.

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*, 812-821.

McFarland, L. A., & Ryan, A. M. (2006). Toward an integrated model of applicant faking behavior. *Journal of Applied Social Psychology, 36*, 979-1016.

McFarland, L. A., Ryan, A. M., & Ellis, A. (2002). Item placement on a personality measure: Effects on faking behaviour and test measurement properties. *Journal of Personality Assessment, 78,* 348-369.

Menghin, S., & Kubinger, K. D. (1996). Zur Legende: "Testpersonen beantworten dem Computer persönliche und intime Fragen offener als einem Testleiter". Ergebnisse eines Experiments [The legend that subjects give more valid answers to private and intimate questions in computer-assisted vs. in paper-and-pencil tests. Experimental results]. *Zeitschrift für Differentielle und Diagnostische Psychologie, 17*, 163-169.

Morgeson, F. P., Campion, M. A., Diboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683-729.

Mueller-Hanson, R. A., Heggestad, E. D., & Thornton, G. C. III (2006). Individual differences in impression management: An exploratory of the psychological process underlying faking. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*], *48*, 288-312.

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.

Ostrom, T.M., Betz, A.L. & Skowronski, J.J. (1992). Cognitive Representation of Bipolar Survey Items. In N. Schwarz & S. Sudman. *Context effects in Social and Psychological Research,* (pp. 297-311). New York: Springer.

Paulhus, D.L. (1991). Measurement and control of response bias. In J.P. Robinson, P.R. Shaver, & L.S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes,* (pp. 17-59). San Diego, CA: Academic Press.

Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality, 66,* 1025-1060.

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. (2005). Cognitive ability and self-reported efficacy of self-presentation predict faking on personality measures. *Journal of Individual Differences, 26*, 194-206.

Pauls, C. A., & Stemmler, G. (2003). Substance and bias in social desirability responding. *Personality and Individual Differences, 35,* 263-275.

Peeters, H., & Lievens, F. (2005). Situational judgement tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement, 65*, 70-89.

Petty, R. E., & Cacioppo J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer-Verlag.

Ramsay, L. J., Schmitt, N., Oswald, F. L., Kim, B. H., & Gillespie, M. A. (2006). The impact of situational context variables on responses to biodata and situational judgement inventory items. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*]*, 48*, 268-287.

Reid-Seiser, H. L., & Fritzsche, B. A. (2001). The usefulness of the NEO PI-R positive presentation management scale for detecting response distortion in employment contexts. *Personality and Individual Differences, 31,* 639-650.

Richman, W. L., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administrated questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*, 754-775.

Robie, C., Brown, D. J., & Beaty, J. C. (2007). Do people fake on personality inventories? A verbal protocol analysis. *Journal of Business and Psychology, 21*, 489-509.

Robie, C., Zickar, M. J., & Schmit, M. J. (2001). Measurement equivalence between applicant and incumbent groups: An IRT analysis of personality scales. *Human Performance, 14*, 187-207.

Robie, C., Curtin, P.J., Foster, T.C., Phillips, H.L. IV, Zbylut, M., & Tetrick, L.E. (2000). The effect of coaching on utility of response latencies in detecting fakers on a personality measure. *Canadian Journal of Behavioral Science, 32,* 226-233.

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.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review, 16*, 155-180.

Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology, 80,* 607-620

Schneider, J.F., & Hübner, R. (1980). Einfluss von Verfälschungsinstruktionen auf die Bearbeitungszeit von Persönlichkeitsfragebögen. [The influence of instructions to fake on the time needed to respond to personality questionnaires] *Zeitschrift für Experimentelle und Angewandte Psychologie, 27,* 565-597.

Scheider-Düker, M., & Schneider, J.F. (1977). Untersuchungen zu Beantwortungsprozess bei psychodiagnostischen Fragebögen. [Analyses of the process of responding to personality questionnaires]. *Zeitschrift für Experimentelle und Angewandte Psychologie, 24,* 282-302.

Schwarz, N., Hippler, H.-J., & Noelle-Neumann, E. (1992). Cognitive model of response-order effects. In N. Schwarz & S. Sudman. *Context effects in Social and Psychological Research,* (pp. 187-201). New York: Springer.

Seiwald, B. B. (2002). Replicability and generalizability of Kubinger's results: Some more studies on faking personality inventories. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*], 44, 17-23.

Smith, T. W. (1992). Thoughts on the nature of context effects. In N. Schwarz & S. Sudman. *Context effects in Social and Psychological Research,* (pp. 163-184). New York: Springer.

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review, 9,* 219-242.

Steinmayr, R., & Kersting, M. (2008) Verfälschbarkeit von Persönlichkeitstests – ein Problem für die soziale Akzeptanz [Fakability of personality tests – a problem for the social acceptance]? In W. Sarges & D. Scheffer (Hrsg.), *Innovative Ansätze der Eignungsdiagnostik [Innovative approaches to qualification assessments*], (pp. 31-40). Göttingen: Hogrefe.

Strack, F. (1992). „Order effects" in survey research: Activation and information functions of preceding questions. In N. Schwarz & S. Sudman. *Context effects in Social and Psychological Research,* (pp. 23-34). New York: Springer.

Stumpf, H., & Steinhart, I. (1981). *Zur Anfälligkeit der Skalenwerte der deutschen "Personality Research Form" (KA) gegenüber tendenziöser Beantwortung* [*On the susceptibility of the scales in the German „Personality Research Form „ (KA) towards tendentious response behaviour*]. Wehrpsychologische Untersuchungen, Heft 3.

ter Laak, J., van Leuven, M., & Brugman, G. (2000). The effect of the accountability instruction and two job types on the big five scores. *European Journal of Psychological Assessment, 16*, 209-213.

Thumin, F. J., & Barclay, A. G. (1993). Faking behaviour and gender differences on a new personality research instrument. *Consulting Psychology Journal, 45*, 11-22.

Tourangeau, R. (1992). Context effects on responses to attitude questions: Attitudes as memory structures. In N. Schwarz & S. Sudman. *Context effects in Social and Psychological Research,* (pp. 35-47). New York: Springer.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103,* 299-314.

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.

Winkelspecht, C., Lewis, P., & Thomas, A. (2006). Potential effects of faking on the NEO-PI-R: Willingness and ability to fake changes, who gets hired in simulated selection decisions. *Journal of Business and Psychology, 21*, 243-259.

Wright, S. S. & Miederhoff, P. A. (1999). Selecting students with personal characteristics relevant to pharmaceutical care. American Journal of Pharmaceutical Education, 63, 132-138.

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. (1999). Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology, 84*, 551-563.

Ziegler, M., Schmidt-Atzert, L., Bühner, M. & Krumm, S. (2007). Fakability of different measurement methods for achievement motivation: questionnaire, semi-projective, and objective. *Psychology Science* [latter: *Psychology Science Quarterly*]*, 49*, 291-307.

5. Original Papers of the Doctoral Thesis

(Paper 1 – 3)

5.1. Paper 1

Khorramdel, L., & Kubinger, K. D. (2006). The effect of speediness on personality questionnaires: an experiment on applicants within a job recruiting procedure. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*], *48*, 378-397.

Psychology Science, Volume 48, 2006 (3), p. 378-397

The effect of speediness on personality questionnaires: an experiment on applicants within a job recruiting procedure

LALE KHORRAMDEL[1] & KLAUS D. KUBINGER

Abstract

The authors conducted an experiment to determine how a particular design of personality questionnaires influences applicant responses on personality scales. A completely crossed 2 x 2 x 2 design was carried out with real-world applicants and individuals in a job application training program in which speed (with or without a time limit), response format (dichotomous or analogue), and instructions (neutral standard instruction or a repeated warning that people who fake can be detected) were manipulated. Two hundred eight participants completed the Myers-Briggs Type Inventory and a German Interpersonal Circumplex (IPC)-based questionnaire. Although providing a warning showed no influence, response format and the interaction between speed and response format showed a significant effect for some scales.

Key words: personality questionnaire, faking good, social desirability, personnel selection, psychological assessment, response format, instruction, speed

[1] Lale Khorramdel and Klaus D. Kubinger, Center of Testing and Consulting, Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna.
Correspondence concerning this article may be addressed to Lale Khorramdel, University of Vienna, Liebiggasse 5, 1010 Vienna, Austria, Email : lale.khorramdel@univie.ac.at, or Klaus Kubinger, Email: klaus.kubinger@univie.ac.at

Personality questionnaires are the best known and the most popular tools used to measure personality. However, personality questionnaires often show a high transparency; that is, it is often evident to the test-taker what constructs the test measures. Because test-takers can infer what constructs items may measure, they may distort their responses in order to present themselves favourably. This may be particularly problematic in the context of personnel selection, where applicants may "fake good" in an attempt to secure a job offer (cp. Kanning & Holling, 2001; Karner, 1999, 2002).

Considerable research has shown that even voluntary participants are able to intentionally fake good when instructed to empathize with a selection candidate (Kubinger, 1996; 2002) or to adapt to a given job profile (Hoeth, Büttel, & Feyerabend, 1967; Lammers & Frankenfeld, 1999). Krahé and Hermann (2003) found similar results when analysing the susceptibility of the NEO-Five Factor Inventory (NEO-FFI) to systematic response tendencies. Because of these potential faking effects, data from self-descriptions should always be regarded carefully (Deller & Kuehn, 2003).

Faking tendencies in real-world selection situations, however, are actually fewer than in simulated situations. Some studies show that adjusting personality scores based on social desirability scores does not decrease the validity of a test (Hough et al., 1990; Moorman & Podsakoff, 1992; Ones, Viswesvaran & Reiss, 1996; Ones, Viswesvaran & Schmidt, 1993), and there is even an established opinion that personality questionnaires are valid methods for personnel selection despite their high transparency (Schmidt & Hunter, 1998; cf. also Marcus, 2003). However, the extent to which validity is decreased by the influence of social desirability bias is unknown (Kanning, 2003). Furthermore, because candidates who fake are more likely to be selected than those who answer honestly, faking may make selection systems unfair (Ellingson, Sackett & Hough, 1999; Hough, 1998). Therefore, test-users should take precautions to prevent or reduce applicant faking on personality questionnaires (Hough & Ones, 2002; McFarland, 2003).

Past research has explored whether it is possible to detect individuals who may be faking. Two means of detection have primarily been used: measuring/analysing response latencies (i.e., the time between item responses; Esser & Schneider, 1998; Holden & Hibbs, 1995; Holden, Kroner, Fekken & Popham, 1992; Hsu, Santelli & Hsu, 1989; Kuntz, 1974; Robie et al., 2000; Schneider & Hübner, 1980) and imbedding social desirability scales (a.k.a., lie scales) within personality measures (Crowne & Marlowe, 1960; Edwards, 1957; Hoeth, Büttel & Feyerabend, 1967; Paulhus, 1991; Schneider-Düker & Schneider, 1977). In the detection literature using response latencies, the general assumption is that response latencies indicate the fidelity of the response. Response latencies may indicate whether a participant's response reflects their self-concept (i.e., an honest response) or a response style (i.e., a faked response). In addition, response latencies may indicate that a test-taker has responded at random (which would affect the reliability and validity of a score; cp. Wagner-Menghin, 2002). Holden and colleagues (Holden & Hibbs, 1995; Holden, Kroner, Fekken, & Popham,1992) proposed a model of personality test item response dissimulation. In their model, a respondent attempts to compare test item content to either a relevant cognitive self-schema or to an adopted schema (for example, faking good represents an adopted schema). They found that responses congruent with self-schemas are faster than when answers are not congruent with self-schemas. These results suggest that adopting the schema to fake good may produce longer response latencies, which can be used to identify fakers. However, the authors acknowledge a number of limitations in their research. First, their research has fo-

cused on dimensions of maladjustment rather than personality scales. Second, longer response latencies are associated with items that have relatively extreme social desirability levels, have extreme endorsement proportions, and are predominately positively-keyed (rather than reverse-keyed). Third, their design compared the response latencies of volunteer participants instructed to fake versus to volunteer participants instructed to answer honestly. Indeed, the results of other response latency studies have produced discrepant results. Kuntz (1974) found significantly longer latencies under both "fake bad" and "fake good" instructions than under standard conditions. In contrast, Hsu, Santelli and Hsu (1989) found shorter latencies under both faking conditions.

The benefits of social desirability (lie-) scales or repeated items to check for consistency (control items) are debatable, mostly, because of their high transparency (Seiwald,2003). Even control items, if recognized, decrease the participant's motivation to answer honestly. Moreover, the validity of lie scales seems doubtful, because they measure not only the participant's tendency to fake but also a personality trait. Attempts to use social desirability scales to statistically adjust personality scores decreases rather than increases the criterion-related validity of the personality measures (Borkenau & Ostendorf, 1992; McCrae & Costa, 1983; Ones et al., 1996; Piedmont et al., 2000; cf. also Hülsheger et al., 2004). Hülsheger et al. (2004) argue that even the attempt to identify and exclude invalid profiles with these scales is a poor strategy because honestly-responding participants may also be erroneously eliminated. Test-takers who have high social desirability scores are not necessarily faking; indeed, high-scorers might simply have answered the questionnaire honestly but have a high degree of the trait measured by social desirability scales. Thus, correcting questionnaires for faking is particularly concerning given that such corrections may have considerable influence on who receives a job offer, yet there is an absence of empirical evidence to support the use of these corrections (Goffin & Christiansen, 2003).

Efforts to suppress faking good

Recent attempts have aimed at making personality questionnaires less fakeable by adjusting aspects of how they are administered. These attempts include adjusting the response format, method of administration, and item positioning. In a summary of this research, no general conclusions could be drawn for these adjustments as the effects appear to be influenced by many moderating variables (Kubinger, 2003a). However, some research suggests that using analogue scales (in which participants mark along a continuous line to indicate the extent of their agreement) as a response format may be less prone to faking than a dichotomous response, multiple-choice, or Q-Sort format (cf. Seiwald, 2002). Questionnaires administered with either paper and pencil or with a computer have not shown any difference in fakeability, nor have verbal as opposed to non-verbal questionnaires (cf. Amelang, Schäfer and Yousfi, 2002). With respect to the effects of the item-positions, it has been shown that test-takers are more likely to fake their answers at the beginning rather than at the end of a questionnaire.

In addition, researchers have attempted to limit faking by adjusting the instructions given to test-takers (Mummendey, 1999). Typically, personality measures are administered with the instruction to answer "as candidly and honestly as possible." A warning instruction goes beyond this by informing test-takers that the test administrator can detect intentional re-

sponse distortion. Hülsheger et al. (2004) found no effect of the warning that "untruthful" response patterns can be detected; however, their sample consisted only of student volunteers. Most research, though, has suggested that these warnings are effective at reducing the prevalence of faking, although the effects are weak (Hoeth & Köbler, 1967; Braun & La Faro, 1968; Dwight & Donovan, 2003). In particular, warning applicants that faking can be detected with an imbedded social desirability scale or by analyzing the response latencies decreases faking (Doll, 1971; Kluger & Colella, 1993; Nias, 1972; Robie et al., 2000; Wheeler, Hamill & Tippins,1996). However, McFarland (2003) calls attention to certain practical consequences of this procedure. Warning the applicants in personality questionnaires tends to provoke negative reactions by the applicants (e.g., Rosse, Miller & Stecher, 1994; Smither et al., 1993; Steiner & Gilliland, 1996). Hence, such a warning may make applicants feel that the employer distrusts them or that applicants cannot present themselves as they would like to be seen. On one hand, this negative reaction may make the most qualified applicants self-select themselves out of the selection process (cf. Ryan, Sacco, McFarland & Kriska, 2000). On the other hand, the applicants' test-taking motivation could be decreased. Hence, such test perceptions may affect selection decisions (Chan et al., 1997), and the validity of the questionnaire (Schmit & Ryan, 1992). Overall, McFarland did not find any negative reactions affected by warnings in her study; however, her sample was limited to voluntary participants who were instructed to imagine a job-application situation.

An indirect method to decrease faking is to exert a time pressure on a participant. Some authors assert that test-takers need more time in order to fake, so adding time pressure may decrease faking tendencies. For example, Bartley (1958) suggested that overly long reaction times are caused by a test-taker searching for substitute responses that will mask his/her initial reaction. Answering without manipulating one's own attitudes takes less time than reflecting some prototypic attitudes in responses. Indeed, the most prior studies support the idea faking causes longer response times; but these have used inadequate samples (such as volunteers rather than job applicants). On the other hand, in an early study, Sutherland and Spilka (1964) have demonstrated that time pressure can result in responses in the direction of what is socially approved. However, they used voluntary students as participants, as well, and the decision for a response interval of 2 seconds per item is neither explained nor evident. The results of Krämer and Schneider (1987) are similar, but, instead of using real time pressure, the participants were only given the instruction to answer quickly and spontaneously. Again, only volunteers served as participants, and, furthermore, the sample was considerably small. Neubauer and Malle (1997) likewise use a speed instruction, and the results show a lower mean neuroticism score in the Eysenck-Personality-Inventory (EPI). Again, however, the study is based on volunteers, and there was no real time pressure.

Aim of the current study

Although the results of some previous studies aiming to reduce applicant faking have been encouraging (particularly the analogue scale), the majority have primarily used non-applicant volunteer samples. The present experiment uses real-world applicants to test the effects on faking of manipulating speed, response format, and instruction.

Hypotheses

The first hypothesis is that a limited response time may decrease the phenomenon of faking good in real-world selection situations. We expect this decrease in faking because the time pressure will preclude the test-taker from thinking about the best (most socially desirable) answer and force the test-taker to answer spontaneously. The second hypothesis is that an analogue scale response format might lead to a more honest self-presentation than would a dichotomous one. We expect that an analogue scale is more difficult to fake because it is more difficult to determine what intensity is socially desired but not suspect of indicating a faked response. The third hypothesis is that a warning instruction may lead to less socially desirable answers.

To detect possible fakers within the scales of the personality questionnaire used, we assume that it is more advantageous for a participant in selection situations to show high values for certain scales and low values for other scales. We expect that fakers would inflate their scores on the scales Self-Concept of Own Competences, Internality, Extroversion, and Thinking, whereas we expect that fakers would provide lower ratings for the scales Powerful Others Control, Chance Control, and Introversion (a description of the personality tests used is given below). Test-takers with such scores might try to present themselves in a socially desirable manner. These predictions are guided by what is commonly believed to be desirable in the jobs for which the applicants in our sample applied (office managers, salesmen/-women, tradesmen/-women, and middle echelon managers). For the scales Feeling, Judging, Perceiving, Intuition, and Sensing, it is not clear whether high or low values would indicate faking because the scales are not clearly indicative of desirable traits for these specific jobs.

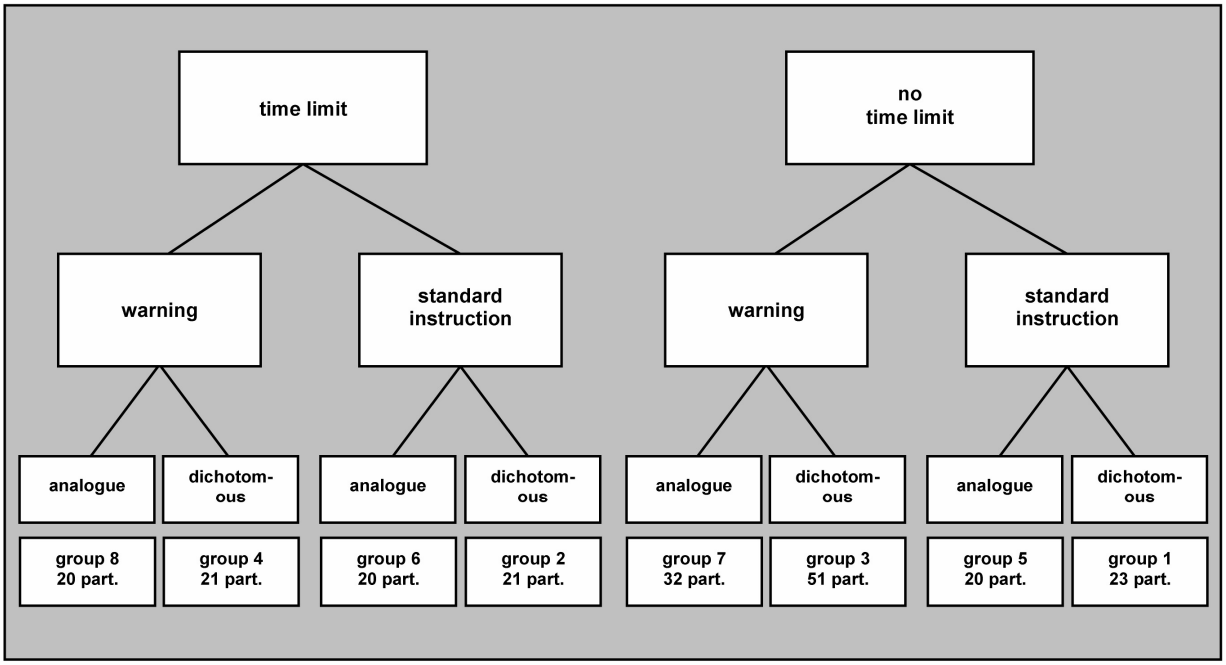
Method

Design

To test whether time limitations, scale response format, and warning instructions affect applicant faking, participants completed personality measures in a completely crossed 2 x 2 x 2 design. Thus, participants were randomly assigned to one of eight groups, representing a combination of the three factors, as shown in Figure 1. Figure 1 also provides the information about the sample size in each of the eight groups. Manipulations of each independent variable are described below.

Response Time and Speed. As is often the case with ability and achievement measures, we constructed a speeded questionnaire. The participants either received an overall limited response time for the items (per page of items) or they received no time limit. The time limitations were based on the results of a pilot test with 10 participants between the age of 18 and 56 who came from different educational backgrounds (primary and secondary education). Each of the 10 participants completed the personality questionnaires as quickly as possible. The time per page each participant needed was recorded, and the means were calculated. Thus, we chose time limits between 45 seconds and 1 minute and 40 seconds for each page of the questionnaire (depending on the number of items per page). Three filler items were presented at the end of each page, which served to guarantee that the time limit

Figure 1:
Experimental design



Note: part. = participants

did not preclude slower test-takers from completing all relevant personality items. Test-takers were instructed to start with the first item on each page, and not to leave a single item out.

Item Response Format. As in previous experiments, participants completed personality questionnaires with either a dichotomous response format or an analogue scale response format. That is, participants with a dichotomous response format decided between total agreement to the given statement of an item or total disagreement ("yes" or "no," and "true" or "false"); participants with the analogue scale response format indicated their response by making a mark on a continuous line. There were at least 39 invisible points on a line in accordance with the length of the 39 *mm* line, a length based on the layout of the answer sheet. However, the analogue scale was scored dichotomously, so that marks on the left half of the line indicated "true/agree" and marks on the right half indicated "false/disagree."

Instruction. All participants received a conventional, neutral instruction. Those in the warning group also received an additional warning that faking can be detected. The warning was given once at the beginning of the questionnaire and once again in the middle of the questionnaire to ensure that the participants did not forget.The standard instruction was:

> There are no correct or incorrect answers in this questionnaire. Your answers merely provide information on how you see things or how you normally make a decision.

The additional warning instruction was:

> Afterwards your answers will be checked by a complex computer-based evaluation programme, in order to ascertain whether your answers are given in an honest manner. Therefore, it does not pay off to fake the questions. You would then simply be asked to answer the questionnaire again.

Measures

It seemed important that the questionnaire be neither too short (to avoid giving the impression that little effort was required), nor too long (to avoid making participants fatigued or frustrated). The final questionnaire set was a battery of well-known paper-pencil personality questionnaires.

Myers-Briggs Type Indicator (MBTI) – German edition. The MBTI (Bents & Blank, 1991) is based on Jung´s personality typology and consists of 90 items. The scales are: Extroversion, Introversion; Sensing, Intuition; Thinking, Feeling; Judging, Perceiving. Descriptions of these scales are provided in Table 1. Because the fifth item of the MBTI originally has three response categories, it was necessary to remove the middle category to create a dichotomous response format for this item. Reliability coefficients were calculated for each scale and range from .007 to .639 (see Table 2).[2]

Interpersonal Circumplex (IPC)-based questionnaire. The German IPC-based questionnaire (FKK; Krampen, 1991) consists of 32 items that measure the "locus of control of reinforcement" concept from J.B. Rotter´s social learning theory of personality (Rotter, 1982; Rotter, Chance & Phares, 1972; Rotter, Seeman & Liverant, 1962). The scales are: Self-Concept of Own Competences, Internality, Powerful Others Control, and Chance Control. Descriptions of these scales are provided in Table 1. The items from the FKK also originally had six response categories; therefore, only the categories "right" and "wrong" were used. The split-half reliabilities of the FKK scales presented in its test-manual range from .63 to .79 (see Table 3).

As explained before, we added three additional items at the end of each page of the questionnaire – altogether 42 items – to guarantee that almost all interesting items (the items of the MBTI and the FKK) are actually answered by each participant, despite having a time limit. These filler items were not analysed. In addition, neither measure originally had an analogue scale. Therefore, we had to establish one. Altogether, the resulting questionnaire consisted of 164 items, of which 122 were actually analysed (omitting the 42 filler items), This questionnaire was presented either with a dichotomous response format or an analogue scale.

---

[2] We thank an anonymous reviewer for suggesting that we note the links between all the scales and the Big Five dimensions of personality. Because there is no empirical evidence of such links between the FKK and the Big Five dimensions, we simply give some plausible correspondence in a separate column. However, the links between the MBTI and the Big Five dimensions are based on the findings of McCrae and Costa (1989), which were supported by the findings of Furnham, Moutafi and Crump (2003).

Table 1:
Description of the scales of the personality questionnaires

Scales	Description	Correspondence to the Big Five dimensions of personality (McCrae & Costa, 1989; Furnham, Moutafi, & Crump, 2003)
MBTI:		
extroversion	external orientation, extrovert attitude	extroversion (E)
introversion	internal orientation, introvert attitude	extroversion (E)
sensing	sensual perception; perceptual processes by means of the five senses; orientation on experiences in the present (here and now)	openness to experience (O)
intuition	intuitive perception; perception of possibilities, meanings and relations which happens by insight	openness to experience (O)
thinking	analytical judgment; judgment due to logical linked imaginations	agreeableness (A)
feeling	emotional judgment; judgment due to personal and social values	agreeableness (A)
judging	judging attitude; focus on decisions and planning of action sequences	conscientiousness (C)
perceiving	perceptual attitude; focus on receipt and perception of information	conscientiousness (C)
German IPC-based questionnaire (FKK):		Plausible correspondence to the Big Five dimensions of personality
self-concept of own competences	generalized expectation to have action possibilities – at least one – at disposal in life of action situations	neuroticism (N)
internality	subjectively noticed control of own life and events of the person specific environment	openness to experience (O)
powerful others control	generalized expectation that important events in life depend on the influence of others	agreeableness (A)
chance control	generalized expectation that life and important events in it depend on destiny, fortune, bad luck and chance	neuroticism (N)

Table 2:
Cronbachs alpha and split-half reliability (Spearman-Brown) of the MBTI-scales (from the current data; n = 208)

MBTI-scales	*Cronbachs Alpha*	*Spearman-Brown*
extroversion	.268	.244
introversion	.322	.348
sensing	.274	.387
intuition	.007	.188
thinking	.074	.040
feeling	.047	.146
judging	.631	.639
perceiving	.639	.608

Table 3:
Split-half reliability (Spearman-Brown) of the FKK-scales (Krampen, 1991)

Study	N	self-concept of own competences	internality	powerful others control	chance control
1	62	.79	.74	.72	.73
2	258	.70	.67	.75	.78
3	152	.72	.63	.65	.67
4	38	.71	.64	.70	.76
5	248	.72	.68	.70	.69
6	2028	.71	.64	.67	.70

Sample

Two hundred eight participants completed the personality questionnaire. Initially, we intended to use only test-takers who were actual job applicants being recruited. Participant data was gathered from two separate sources. First, 113 of the participants were recruited from a special job-application training course consisting of long-term unemployed individuals within a re-education programme. Near the end of this programme, participants were assessed as part of the course training to prepare for real-world job-applications. That is, for this group of participants, the questionnaire was administered as a simulated-selection process. This testing was part of an evaluation of the effects of the training programme. The participants received personal feedback regarding their individual results. Second, 95 participants were real-world job applicants whose data were taken from two personnel and management consulting companies.

Altogether, 96 women and 112 between the age of 18 and 56 with various educational backgrounds were tested by seven test instructors. These instructors had received an exact verbal instruction and written guide describing how to instruct the participants. The test-takers were randomly assigned to the eight experimental conditions; however, their sex and original institution were noted to ensure that men, women, and participants of each of the three institutions are represented adequately in each of the 8 experimental groups.

Because data collection occurred as an on-going process at different locations, it was not possible to assign exactly the same number of participants to each group. That is, we had to conform with the routine of the different institutions. As a result, 116 participants filled out the questionnaire with the dichotomous response format, and 92 participants filled out the questionnaire with the analogue scale response format. Eighty-two participants were given the questionnaire without any time limit, and 126 participants were given the questionnaire with the time limit described above. Eighty-four participants were given only the standard instruction, and 124 were additionally given the warning instruction. The participants of the special training course were tested in groups (with a maximum of 15 persons per group), whereas the job applicant sample was tested individually.

Results

The means and standard deviations for all scales in each experimental condition are given in Table 4. In addition, Appendix A shows the intercorrelations between all scales for the given data.

A multivariate analysis of variance (MANOVA) was used to compare means across conditions (α = .05). With a sample size of 208/8=26 and α = .05, a MANOVA has adequate power (.80) to detect a mean difference of 2/3 standard deviations (Rasch & Kubinger, 2006). To test the homogeneity of variance across cells, a Levene's test was calculated for each scale. However, four MBTI scales (Extroversion, Introversion, Thinking, and Feeling) failed the Levene's test for homogeneity of variance (p = .047, .006, .018, and .013, respectively). Hence, only the scales Judging, Perceiving, Intuition, and Sensing from the MBTI and the scales Self-Concept of Own Competences, Internality, Powerful Others Control, and Chance Control from the FKK are taken into consideration in the following. Box's M-Test for testing the homogeneity of the variance-covariance matrix was conducted on the remaining scales but was not significant (p = .239), indicating that the resulting F-values of multivariate analysis of variance may be interpreted. Table 5 shows the results of the MANOVA testing the main and the interaction effects of the three factors. The multivariate analysis of variance shows a significant effect of the response format (dichotomous vs. analogue), and a significant interaction effect between response format and time (limited time vs. unlimited time).

To more clearly understand these significant effects, each scale was considered individually. Tables 6 and 7 present the result of univariate factorial ANOVAs, the associated effect sizes, and their confidence intervals. Either a (significantly) higher or lower score (dependent on the meaning of the particular scales) may be interpreted as indicating faking good.

Table 6 and 7 show that only one scale (Internality) showed significantly different means across response format groups (p = .038). However, for the time x response format interaction, Self-Concept of Own Competences (p < .001), Powerful Others Control" (p = .008), and Chance Control (p < .001) showed significant interaction effects. To determine the direction of significant differences, we examined the means of the scores (see Tables 8 and 9).

Table 4:
Means and standard deviations (SD) for all scales by experimental condition
(8 experimental groups)

	No Warning				Warning			
	Dichotomous		Analogue		Dichotomous		Analogue	
Scale	No Lim.	Lim.	No Lim.	Lim.	No Lim.	Lim.	No Lim.	Lim.
extroversion	16.77	14.86	13.15	16.30	15.41	15.43	15.66	14.68
	(5.84)	(6.03)	(4.40)	(5.24)	(4.55)	(7.26)	(5.15)	(6.58)
introversion	7.82	9.67	11.55	8.55	9.29	9.43	8.88	9.68
	(5.71)	(5.65)	(4.15)	(4.89)	(4.36)	(6.79)	(4.97)	(6.57)
sensing	10.05	10.05	12.50	11.40	11.65	9.95	10.88	10.68
	(3.51)	(4.82)	(4.03)	(3.17)	(4.25)	(4.91)	(4.08)	(5.94)
intuition	7.14	7.57	6.50	6.35	6.33	7.67	7.25	7.58
	(2.98)	(3.63)	(2.96)	(2.56)	(3.06)	(3.92)	(3.37)	(4.43)
thinking	10.45	9.57	10.00	11.05	9.43	9.43	8.53	9.47
	(2.92)	(4.37)	(2.25)	(2.50)	(3.53)	(2.68)	(3.11)	(3.01)
feeling	6.95	7.52	7.25	6.10	7.61	7.86	8.47	8.11
	(2.57)	(3.50)	(1.92)	(2.36)	(3.02)	(2.13)	(2.86)	(2.47)
judging	10.73	11.19	10.55	10.70	10.63	10.19	10.00	10.89
	(2.99)	(3.46)	(2.82)	(2.56)	(3.30)	(3.68)	(3.21)	(3.43)
perceiving	6.14	6.24	7.10	6.15	6.75	7.10	7.22	6.32
	(3.96)	(3.83)	(3.67)	(3.57)	(4.02)	(4.33)	(4.29)	(4.22)
self-concept of own comp.	1.18	2.76	2.65	1.40	2.25	3.05	2.75	1.68
	(1.87)	(2.43)	(1.81)	(1.27)	(2.07)	(3.13)	(1.97)	(1.70)

Note. Within each row, means are presented above and standard deviations presented below in parentheses.

The additional warning instruction not to fake did not influence the level of the score (see Table 5). However, the kind of response format does have an influence (see Table 6), although only for a single scale, Internality. Table 8 shows that if the dichotomous response format is used, a slightly higher tendency towards Internality is exhibited than when an analogue scale response format.

Table 5:
Multivariate Analysis of Variance – including the scales "judging", "perceiving", "intuition", and "sensing" from the MBTI and the scales "self-concept of own competences", "internality", "powerful others control", and "chance control" from the IPC-like questionnaire (FKK)

Effect		Value	*F*	Hypothesis *df*	Error *df*	Sig.	Partial Eta Squared
format	Pillai's Trace	.088	2.307	8	191	.022	.088
instruction	Pillai's Trace	.022	.526	8	191	.836	
time	Pillai's Trace	.019	.465	8	191	.880	
format * instruction	Pillai's Trace	.060	1.528	8	191	.150	
format * time	Pillai's Trace	.181	5.277	8	191	.000	.181
instruction * time	Pillai's Trace	.044	1.095	8	191	.368	
format * instruction * time	Pillai's Trace	.071	1.826	8	191	.074	

Table 6:
Scale-wise F-values with respect to the factor Format

Dependent Variable	*df*	*F*	Sig.	Effect Size	Confidence-	Interval
					lower bound	*upper bound*
judging	1	.098	.755			
perceiving	1	.058	.809			
intuition	1	.274	.601			
sensing	1	2.171	.142			
self-concept of own competences	1	.439	.508			
internality	1	4.343	.038	.3444	.0674	.6214
powerful others control	1	2.845	.093			
chance control	1	.600	.440			

Table 7:
Scale-wise F-values with respect to the factor Format x Time

Dependent Variable	*df*	*F*	Sig.	Effect Size	Confidence-	Interval
					lower bound	*upper bound*
judging	1	.291	.590			
perceiving	1	.956	.329			
intuition	1	.654	.420			
sensing	1	.025	.875			
self-concept of own competences	1	16.614	.000	.2191	-.1707	.6089
internality	1	.195	.660			
powerful others control	1	7.120	.008	.3402	-.0693	.7497
chance control	1	33.696	.000	.0955	-.3114	.5023

Table 8:
Means of scores at the different levels of the significant factor "Format"

Dependent Variable	Format	Mean
internality	dichotomous	2.297
	analogue scale	1.853

Table 9:
Means of scores at the different levels of the significant interactions of factor "Format" and factor "Time"

Dependent Variable	Format	Time	Mean
self-concept of own competences	dichotomous	no time limit	1.718
		time limit	2.905
	analogue scale	no time limit	2.700
		time limit	1.542
powerful others control	dichotomous	no time limit	5.901
		time limit	5.333
	analogue scale	no time limit	4.659
		time limit	5.613
chance control	dichotomous	no time limit	5.921
		time limit	4.333
	analogue scale	no time limit	4.009
		time limit	5.795

Interpretation

Furthermore, three of the eight scales show a significant interaction effect between response format and time pressure (see Table 7). For the scales Powerful Others Control and Chance Control, the dichotomous response format produced a higher mean when paired with no time limit, whereas the analogue scale response format produced a higher mean when paired with a time limit. On the contrary, the scale Self-Concept of Own Competences, showed the opposite pattern (higher means were observed when the dichotomous format was paired with a time limit and when the analogue format was paired with no time limit). Supposing the participants think of a high degree of Self-Concept of Own Competences as being highly socially desirable and being low on convincement of external control, these time x response format effects point in the same direction: the answers to personality questionnaires are faked good if there is either a limited administration time and a dichotomous response format or if there is no limited administration time and an analogue scale response format. That is, participants were more likely to give a socially desirable answer if the time was limited on a dichotomous scale, and they were also able to answer in a more socially desirable manner if they had plenty of time to respond to an analogue scale.

Again, there is additionally an occasional trend of answering in a socially desirable manner if a dichotomous response format is used (see Table 8); this is true in so far as high In-

ternality is socially desirable. However, there is no main effect for time pressure (see Table 5).

Discussion

The current experiment used applicants from a job recruiting procedure to examine the effects of time pressure, response format, and warnings on faking. Forty-six percent of the sample consists of such job applicants, and 54% of the participants came from a special job-application training course. Of past research examining the effects of such administrative adjustments on personality questionnaire faking, very few studies have been designed like our experiment. That is to say, the current study does not apply a faking good instruction and test the respective effect only on volunteers. Nor is it an experiment that considers different administration conditions using only volunteers. Instead, the current study examines how these adjustments affect personality scale responses by actual job applicants. Furthermore, two factors of potential inflationary influence on item responses (good) are considered here: the influence of a warning instruction on the one hand, and the influence of a speeded administration on the other hand. Finally, the benefit of using an analogue scale response format was investigated once again, because evaluations so far have not disclosed any unequivocal results.

Although the observed results do not provide a clear and consistent method of administering personality measures that prevents applicant faking, the results do suggest several conclusions. There is some evidence that an analogue scale response format tends to be superior to a dichotomous response format if the psychologist is aware that the faking good phenomenon might occur. However, using an analogue scale response format is not in any way a guarantee for preventing faking good; this response format probably works in only a few of the conceivable personality questionnaires' scales. Furthermore, the response format effect of the analogue scale might be enhanced by imposing a time limit for answering the questionnaire items, whereas the dichotomous response format shows the same effect without imposing a time limit. Indeed, three of the eight scales tested produced such an interaction effect (cp. Table 7).

Again, no means have been discovered that prevent faking good in any case of personality scales. Unfortunately, warning applicants that faking can be detected did not work at all. Future research should investigate the effect of imposing a time limit for each individual item rather than for a set of items. This manipulation may make future results even more pronounced than those observed here. In addition, future research should identify a specific class of personality dimensions and personality questionnaires for which the analogue scale response format works to prevent faking good.

The current study has several limitations that future research should address. First, our sample was not a homogeneous set of job applicants; rather it contained both real job applicants and unemployed individuals participating in a job application training programme that simulated applying for a job. However, the latter group did actually carry out a job-application and had the possibility of experiencing how they would have performed in a real-world selection situation. Furthermore, these individuals had recently been through a similar selection procedure, making a real applicant setting salient. Thus, these factors suggest that these individuals completed the personality measure in a setting quite similar to real-world

job candidates. Nonetheless, there is still the possibility that they may not have had the motivation to distort their responses, or their distortions may be based on generic social desirability of traits rather than traits relevant for a particular job (as would be the case for actual applicants).

Second, it is unclear whether limiting response time (speed testing) changes the constructs being measured by a personality questionnaire. For example, such time limitations may mean the questionnaire also measures the ability to work under pressure, the ability to cope with stress, the motivation to deliver exceptional performance (i.e. trying to be as fast as possible), or the ability to understand the contents of a question quickly. Hence, certain participants may have been handicapped--namely, those who would need more time to understand the meaning of a question and who therefore may have not been able to answer in a manner representing their real behaviour or attitudes. In this case, individuals may have misunderstood items or answered at random. Thus, future studies should also conduct pretests of verbal comprehension and examine the baseline reaction times.

References

Allport, F.H. (1920). The influence of the group upon association and thought. Journal of Experimental Psychology, 3, 159-182.

Amelang, M., Schäfer, A., & Yousfi, S. (2002). Comparing verbal and non-verbal personality scales: Investigating the reliability and validity, the influence of social desirability, and the effects of fake good instructions. Psychologische Beiträge [latter: Psychology Science], 44, 24-41.

Bartley, S.H. (1958). Principles of perception. New York: Harper & Brothers.

Bents, R., & Blank, R. (1991). Der Myers-Briggs Typenindikator (MBTI) [Myers- Briggs Type Indicator]. Weinheim: Beltz.

Borkenau, P., & Ostendorf, F. (1992). Social desirability scales as moderator and suppressor variables. European Journal of Personality, 6, 199-214.

Braun, J.R., & La Faro, D. (1968). Effects of salesman faking instructions on the Contact Personality Factor Test. Psychological Reports, 22, 1245-1248.

Brehm, J.W. (1966). A theory of psychological reactance. New York: Academic Press.

Chan, D., Schmitt, N., DeShon, R.P., Clause, C.S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. Journal of Applied Psychology, 82, 300-310.

Crowne, D.P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. Journal of Consulting Psychology, 24, 349-354.

Deller, J., & Kuehn, M. (2993). Occupational Personality Questionnaire (opq32). In E. Fay (Hg.), Tests unter der Lupe 4: Aktuelle psychologische Testverfahren – kritisch betrachtet (pp. 76-104), Göttingen: Vandenhoeck & Ruprecht.

Doll, R.E. (1971). Item susceptibility to attempted faking as related to item characteristics and adopted fake set. Journal of Psychology, 77, 9-16.

Dwight, S.A., & Donovan, J.J. (2003). Do warnings not to fake reduce faking? Human Performance, 16, 1-23.

Edwards, A.L. (1957). The social desirability variable in personality assessment and research. New York, NY: Dryden Press.

Ellingson, J.E., Sackett, P.R., & Hough, L.M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. Journal of Applied Psychology, 84, 155-166.

Esser, C., & Schneider, J.F. (1998). Differentielle Reaktionslatenzzeiten beim Bearbeiten von Persönlichkeitsfragebogen als möglicher Indikator für Verfälschungstendenzen [Differential response latencies by answering personality questionnaires as possible indicator for faking-tendencies]. Zeitschrift für Differentielle und Diagnostische Psychologie, 19, 246-257.

Flavell, J.H., Juris, D., Feinberg, L.D., & Budin, W. (1958). A microgenetic approach to word association. Journal of Abnormal and Social Psychology, 57, 1-7.

Goffin, R.D., & Christiansen, N.D. (2003). Correcting Personality Tests for Faking: A Review of Popular Tests and an Initial Survey of Researchers. International Journal of Selection and Assessment, 11, 340-344.

Hoeth, F., Büttel, R., & Feyerabend, H. (1967). Experimentelle Untersuchungen zur Validität von Persönlichkeitsfragebogen [Experimental investigation on the validity of personality questionnaires]. Psychologische Rundschau, 18, 169-184.

Hoeth, F., & Köbler, V. (1967). Zusatzinstruktion gegen Verfälschungstendenzen bei der Beantwortung von Persönlichkeitsfragebogen [Additional instruction against faking-tendencies on answering personality questionnaires]. Diagnostica, 13, 117-130.

Holden, R., & Hibbs, N. (1995). Increment validity of response latencies for detecting fakers on a personality test. Journal of Research in personality, 3, 362-372.

Holden, R.R., Kroner, D.G., Fekken, G.C., & Popham, S.M. (1992). A model of personality test item response dissimulation. Journal of Personality and Social Psychology, 63, 272-279.

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., & McCloy, R.A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. Journal of Applied Psychology, 75, 581-595.

Hough, L.M., & Ones, D.S. (2002). The structure, measurement, validity, and use of personality variable in industrial, work, and organizational psychology. In N. Anderson, D.S. Ones, H.K. Sinangil & C. Viswesvaran (eds), International handbook of work and organizational psychology. Sage Publications.

Hsu, L.M., Santelli, J., & Hsu, J.R. (1989). Faking detection validity and incremental validity of response latencies to MMPI Subtle and Obvious item. Journal of Personality assessment, 53(2), 278-295.

Hülsheger, U.R., Spinath, F.M., Küppers, A., & Etzel, S. (2004). Experimentelle Untersuchung zweier Methoden zur Reduzierung Sozialer Erwünschtheit in einem computergestützen eignungsdiagnostischen Testverfahren [Experimental investigation of two methods for reduction of social desirability within a computer based diagnostic test of qualification]. Zeitschrift für Personalpsychologie, 3(1), 24-33.

Kanning, U.P. (2003). Sieben Anmerkungen zum Problem der Selbstdarstellung in der Personalauswahl [Seven notes on the Problem of self-projection in personnel selection]. Zeitschrift für Personalpsychologie, 2(4), 193-197.

Kanning, U.P., & Holling, H. (2001). Struktur, Reliabilität und Validität des NEO-FFI in einer Personalauswahlsituation [Structure, reliability and validity of the NEO-FFI in situations of personnel selection]. Zeitschrift für Differentielle und Diagnostische Psychologie. 22, 239-247.

Karner, T. (1993). Eine empirische Anwendung des Modells von Müller für kontinuierliche Antwortskalen (mittels des computerisierten Myers-Briggs Typenindikators) [An empirical

use of the model of Müller for continuous response scales (by means of the computer based Myers-Briggs Type Indicator)]. Unpublished degree dissertation, University of Vienna, Vienna.

Karner, T. (1999). Eine systematische Untersuchung der Auswirkungen verschiedener Antwortmodi auf die Qualität Psychologischer Fragebögen [A systematical investigation of the effect of different response modes on the quality of psychological questionnaires]. Unpublished doctoral dissertation, University of Vienna, Vienna.

Karner, T, (2002). The volunteer effect of answering personality questionnaires. Psychologische Beiträge [latter: Psychology Science], 44, 42-49.

Kluger, A.N., & Colella, A. (1993). Beyond the mean bias: The effect of warning against faking on biodata item variances. Personnel Psychology, 46, 763-780.

Krämer, H.J., & Schneider, J.F. (1987). Validität von Fragebogendaten in Abhängigkeit von Antwortzeit-Instruktion und der intraindividuellen Variabilität der Probanden [Validity of questionnaires dependent on response-instruction and intraindividual variability of participants]. Psychologische Beiträge [latter: Psychology Science], 29, 458-468.

Krahé, B., & Herrmann, J. (2003). Verfälschungstendenzen im NEO-FFI: Eine experimentelle Überprüfung [Fake-tendencies in the NEO-FFI: an experimental investigation]. Zeitschrift für Differentielle und Diagnostische Psychologie, 24, 105-117.

Krampen, G. (1991). Fragebogen zu Kompetenz- und Kontrollüberzeugungen (FKK) [Questionnaire of competence and control persuasion]. Göttingen (Germany): Hogrefe.

Kubinger, K.D. (1996). Zur Leichtgläubigkeit der Psychologen: Die unselige Anwendung von Persönlichkeitsfragebögen [On the credulity of psychologists: The unfortunate use of personality questionnaires]. In M. Jirasko, J. Glück & B. Rollett (Eds.) (1996). Perspektiven psychologischer Forschung in Österreich [Perspectives of psychological research in Austria] (pp. 87-91). Vienna: WUV.

Kubinger, K.D. (2002). On faking personality inventories. Psychologische Beiträge [latter: Psychology Science], 44, 10-16.

Kubinger, K.D. (2003a). (Un-) Verfälschbarkeit. In K.D. Kubinger, & R.S. Jäger (Eds.), Schlüsselbegriffe der Psychologischen Diagnostik [Keynotes to Psychological Assessment] (pp. 429-432). Weinheim (Germany): Beltz.

Kubinger, K.D. (2003b). Objektiver Persönlichkeitstest. In K.D. Kubinger, & R.S. Jäger (Eds.), Schlüsselbegriffe der Psychologischen Diagnostik [Keynotes to Psychological Assessment] (pp. 304-309). Weinheim (Germany): Beltz.

Kuntz, D. (1974). Effects of faking instructions on the word-association test. Psychological Reports, 35, 1183-1192.

Lammers, F., & Frankenfeld, V. (1999). Effekte gezielter Antwortstrategien bei einem Persönlichkeitsfragebogen mit „forced-choice“-Format [Effects of aimed response strategies at a personality questionnaire with forced-choice-format.] Diagnostica, 45, 65-68.

Marcus, B. (2003). Das Wunder sozialer Erwünschtheit in der Personalauswahl. [The miracle of social desirability in personnel selection settings] Zeitschrift für Personalpsychologie, 2(3), 129-132.

McCrae, R.R., & Costa, P.T.Jr. (1983). Social desirability scales: More substance than style. Journal of Consulting and Clinical Psychology, 51, 882-888.

McFarland, L.A. (2003). Warning Against Faking on a Personality Test: Effects on Applicant Reactions and Personality Test Scores. International Journal of Selection and Assessment, 11(4), 265-276.

Moorman, R.H., & Podsakoff, P.M. (1992). A meta-analytic review and empirical test or the potential confounding effects of social desirability response sets in organizational behaviour research. Journal of Occupational and Organizational Psychology, 65, 131-149.

Mummendey, H.D. (1999). Die Fragebogen-Methode [The questionnaire method]. (3rd. Ed.). Göttingen: Hogrefe.

Neubauer, A.C., & Malle, B.F. (1997). Questionnaire Response Latencies: Implications for Personality Assessment and Self-Schema Theory. European Journal of Psychological Assessment, 13, 109-117.

Nias, D.K.B. (1972). The effects of providing a warning about the lie scale in a personality inventory. British Journal of Educational Psychology, 42, 308-312.

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.

Ones, D.S., Viswesvaran, C. &, Schmidt, F.L. (1993). Comprehensive meta-analysis of integrity test validities: findings and implications for personnel selection and theories of job performance. Journal of Applied Psychology, 78, 679-703.

Paulhus, D.L. (1991). Measurement and control of response bias. In J.P. Robinson, P.R. Shaver, & L.S. Wrightsman (Eds.), Measures of personality and social psychological attitudes (pp. 17-59). San Diego, CA: Academic Press.

Piedmont, R.L., McCrae, R.R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. Journal of Personality and Social Psychology, 78, 582-593.

Ponocny, I., & Klauer, K.C. (2002). Towards identification of unscalable personality questionnaire respondents: The use of person fit indices. Psychologische Beiträge [latter: Psychology Science], 44, 94-107.

Rasch, D. & Kubinger, K.D. (2006). Statistik für das Psychologiestudium – Mit Softwareunterstützung zur Planung und Auswertung von Untersuchungen sowie zu sequentiellen Verfahren [Statistics for the study of psychology – software application of planning and sequential procedures]. München: Spektrum.

Robie, C., Curtin, P.J., Foster, T.C., Phillips, H.L. IV, Zbylut, M., & Tetrick, L.E. (2000). The effect of coaching on utility of response latencies in detecting fakers on a personality measure. Canadian Journal of Behavioral Science, 32(4), 226-233.

Rosse, J.G., Miller, J.L., & Stecher, M.D. (1994). A field study of job applicants´ reactions to personality and cognitive ability testing. Journal of Applied Psychology, 79, 987-992.

Rotter, J.B. (1982). The development and application of social learning theory. New York, NY: Praeger.

Rotter, J.B., Chance, J.E., & Phares, E.J. (Eds.) (1972). Applications of a social learning theory of personality. New York, NY: Holt, Rinehart & Winston.

Rotter, J.B., Seeman, M., & Liverant, S. (1962). Internal versus external control of reinforcement: A major variable in behaviour theory. In N.F. Washburn (Ed.), Decisions, values, and groups (Vol. 2, pp. 473-516). London: Pergamon.

Ryan, A.M., Sacco, J.M., McFarland, L.A. &, Kriska, S.D. (2000). Applicant self-selection: Correlates of withdrawal from a multiple hurdle process. Journal of Applied Psychology, 85, 163-179.

Sanders, G.S., & Baron, R.S. (1975). The motivating effects of distraction on task performance. Journal of Personality and Social Psychology, 32, 956-963.

Schmidt, F.L., & Hunter, J.E. (1998). The validity and utility of personnel 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. (1992). Test-taking dispositions: A missing link? Journal of Applied Psychology, 77, 629-637.

Schneider, J.F., & Hübner, R. (1980). Einfluss von Verfälschungsinstruktionen auf die Bearbeitungszeit von Persönlichkeitsfragebögen. [The influence of instruction to fake on time needed to respond to personality questionnaires] Zeitschrift für Experimentelle und Angewandte Psychologie, 27, 565-597.

Scheider-Düker, M., & Schneider, J.F. (1977). Untersuchungen zu Beantwortungsprozess bei psychodiagnostischen Fragebögen. [Analyses of the process of responding to personality questionnaires] Zeitschrift für Experimentelle und Angewandte Psychologie, 24, 282-302.

Seiwald, B.B. (2002). Replicability and generalizability of Kubinger`s results: Some more studies on faking personality inventories. Psychologische Beiträge [latter: Psychology Science], 44, 17-23.

Seiwald, B.B. (2003). Antworttendenzen (response set). In K.D. Kubinger, & R.S. Jäger (Eds.), Schlüsselbegriffe der Psychologischen Diagnostik [Keynotes to Psychological Assessment] (pp. 29-32). Weinheim (Germany): Beltz.

Smither, J.W., Reilly, R.R., Millsap, R.E., Pearlman, K., & Stoffey, R.W.(1993). Applicant reactions to selection procedures. Personnel Psychology, 46, 49-76.

Steiner, D.D., & Gilliland, S.W. (1996). Fairness reactions to personnel selection techniques in France and the United States. Journal of Applied Psychology, 81, 134-141.

Sutherland, B.V., & Spilka, B. (1964). Social desirability, item-response time and item significance. Journal of Consulting Psychology, 28, 447-451.

Usdansky, G., & Chapman L.J. (1960). Schizophrenic-like responses in normal subjects under time pressure. Journal of Abnormal and Social Psychology, 60, 143-146.

Wagner-Menghin, M.M. (2002). Towards the identification of non-scalable personality questionnaire respondents: Taking response time into account. Psychologische Beiträge [latter: Psychology Science], 44, 62-77.

Wheeler, J.K., Hamill, L.S., & Tippins, N.T. (1996). Warning against candidate misrepresentation: Do they work? Paper presented at the 11th annual conference of the Society of Industrial and Organizational Psychology, San Diego.

Appendix see next page.

Appendix

Intercorrelations (Pearson) between all scales (from the current data; n = 208)

	judging	perceiving	feeling	thinking	intuition	sensing	extro-version	intro-version	self-concept of own competences	internality	powerful others control	chance control
judging	1	-.951	-.178	.235	-.490	.480	-.085	.085	-.085	-.234	-.090	.172
perceiving	-.951	1	.143	-.202	.483	-.470	.083	-.085	.107	.264	.073	-.189
feeling	-.178	.143	1	-.943	.342	-.316	-.014	.012	.225	.143	-.099	-.206
thinking	.235	-.202	-.943	1	-.339	.307	.006	-.006	-.227	-.194	.120	.221
intuition	-.490	.483	.342	-.339	1	-.950	.097	-.097	.061	.124	.082	-.102
sensing	.480	-.470	-.316	.307	-.950	1	-.121	.118	-.028	-.155	-.131	.048
extroversion	-.085	.083	-.014	.006	.097	-.121	1	-.983	-.399	-.089	.341	.153
introversion	.085	-.085	.012	-.006	-.097	.118	-.983	1	.424	.107	-.341	-.166
self-concept of own competences	-.085	.107	.225	-.227	.061	-.028	-.399	.424	1	.300	-.442	-.410
internality	-.234	.264	.143	-.194	.124	-.155	-.089	.107	.300	1	-.088	-.192
powerful others control	-.090	.073	-.099	.120	.082	-.131	.341	-.341	-.442	-.088	1	.394
chance control	.172	-.189	-.206	.221	-.102	.048	.153	-.166	-.410	-.192	.394	1

5.2. Paper 2

Khorramdel, L., Kubinger, K. D., & Uitz, A. (submitted). Questionnaire length and impression management: Do applicants just forget to fake? *International Journal of Selection and Assessment*.

Questionnaire length and impression management:

Do applicants just forget to fake?

Lale Khorramdel, Klaus D. Kubinger, and Alexander Uitz
University of Vienna

Lale Khorramdel, Klaus D. Kubinger, and Alexander Uitz Center of Testing and Consulting, Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna.

Correspondence concerning this article may be addressed to Lale Khorramdel, University of Vienna, Liebiggasse 5, 1010 Vienna, Austria, Email : lale.khorramdel@univie.ac.at, or Klaus Kubinger, Email: klaus.kubinger@univie.ac.at

Abstract

An experiment was conducted to investigate the effects of questionnaire length and questionnaire content on socially desirable response behaviour. It was hypothesised that socially desirable response distortion would either increase or decrease towards the end of a very long questionnaire as learning or fatigue effects might occur making it more or less easy to adjust responses to a faking good schema. Furthermore, it was hypothesised that particular questionnaire contents are especially vulnerable to response distortion. Eighty-four applicants filled out a questionnaire consisting of 516 items, whose item positions were varied to test if responses are affected by an overlong test length. Results provide evidence of decreased alertness or concentration towards the end of the questionnaire as a result of fatigue effects.

Key words: faking good, fatigue effects, impression management, item position, psychological assessment, personality questionnaires, personnel selection, questionnaire length, social desirability

Introduction

Intentional response distortion (faking good, impression management) in personality questionnaires in a socially desirable or job-related desirable way is an interesting phenomenon in terms of how response behaviour is affected by different contexts like situation related (e.g. personnel selection), person related (motivation and ability to fake, personality traits) and measure related variables (e.g. content, presentation mode). It is also a frequently discussed and investigated topic (cf. Rothstein & Goffin, 2006) particularly with respect to personnel selection where research shows a lot of controversy.

Whether intentional response distortion should be prevented at any cost as rank order changes take place influencing which applicant gets hired (Ellingson, Sackett, & Hough, 1999; Griffith, Chmielowski, & Yoshita, 2007; Mueller-Hanson, Heggestad, & Thornton, 2003; Robie, Brown, & Beaty, 2007; Rosse, Stecher, Miller, & Levin, 1998; Winkelspecht, Lewis, & Thomas, 2006), or whether it is a form of intelligent adaptation to situations, socially adaptive, or an expression of social competence that could predict job performance (Marcus, 2003a, 2003b; Morgeson, Campion, Diboye, Hollenbeck, Murphy, & Schmitt, 2007) has not been clarified yet and cannot be generalised for all situations and test-takers.

Some researchers argue (mostly with meta-analyses) that construct validity or criterion-related validity of self-reported personality measures are barely affected by intentional response distortion (Bradley & Hauenstein, 2006; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Ellingson, Smith, & Sackett, 2001; Moorman & Podsakoff, 1992; Ones, Dilchert, Viswesvaran, & Judge, 2007; Ones, Viswesvaran & Reiss, 1996; Ones, Viswesvaran & Schmidt, 1993; Smith & Ellingson, 2002). It was also shown that the factor structure of a Greek big five measure remained intact when used in personnel selection (Tsaousis & Nikolaou, 2001). Another study was able to show that faking affected the construct validity but not the criterion validity of a big five measure by modelling faking as a measurement error that is caused by an interaction between context and test-taker

(Ziegler & Bühner, 2009). Moreover, there is the opinion that personality tests can provide substantial criterion validity when used in personnel selection if the measured traits are matched to the nature of the job and are measured precisely enough (Goffin & Boyd, 2009).

Other researchers assume social desirability to be a serious problem and that self-reported personality scales have a lot of shortcomings when used in selection (Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001; Winkelspecht, Lewis, & Thomas, 2006) as criterion-related validity is affected at the high end of the predictor distribution (Mueller-Hanson, Heggestad, & Thornton, 2003). Furthermore, slightly lower reliabilities of questionnaire scales were revealed in applicant samples than in non-applicant samples, and it was shown that the underlying constructs of a questionnaire were measured differently across these samples harming the construct validity (Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). It was shown that hiring decisions are more sensitive to faking than the validity of a questionnaire is (Marcus, 2006). Both depend most importantly on variance in faking, and decision overlap decreased considerably as the variance of faking increased. Given a small selection ration even a small variance of faking can influence the number of decisions significantly.

Strategies to deal with faking good, like the identification of response distortion (with the use of response time latencies or social desirability scales), the discouragement of test-takers to fake (with warning instructions that faking can be identified or will have negative consequences), or efforts to make personality questionnaires less fakable (by adjusting the response format, method of administration, or item positioning) have provided inconsistent results, and no general conclusions can be drawn for these adjustments as the effects may be influenced by many moderating variables (Dilchert, Ones, Viswesvaran, & Deller, 2006).

However, the current paper defines socially desirable response distortion as a context effect moderated by different interacting variables, which is shown by the different models of impression management, whose investigation could provide an important contribution to personality research and social theories.

Models of impression management have been proposed to explain faking behaviour by showing how different moderating variables are linked together and interact with one another. The main key aspects of these models are described in the following. Then context effects which are not considered in these models but which could influence the extent of the faking behaviour are additionally discussed.

Models and theories of impression management

Because some individuals are more capable of faking than others it can be assumed that fakability is a variable of individual differences (Mueller-Hanson, Heggestad, & Thornton, 2006; Viswesvaran, & Ones, 1999), and individuals who have knowledge of the aspired job and the appropriate desired behaviours may be able to fake their responses better (Goffin & Boyd, 2009; Levashina & Campion, 2006; Snell, Sydell, & Lueke, 1999). Faking behaviour also seems to occur with different faking styles, for instance slight versus extreme faking (Robie, Brown, & Beaty, 2007; Zickar, Gibby, & Robie, 2004).

Models, which try to explain the process which underlies the intentional response distortion, describe faking behaviour as an interaction of the ability and motivation to fake (Snell, Sydell, & Lueke, 1999), the opportunity to fake depending on the test-takers true score (McFarland & Ryan, 2000), the perception of the situation (belief in the importance of faking, perceived behavioural control and beliefs about subjective norms), and general personality characteristics like Conscientiousness and Emotional stability (McFarland & Ryan, 2006; Mueller-Hanson, Heggestad, & Thornton, 2006). The theory that cognitive ability (general intelligence) is related to the amount of faking, to the ability to perceive the

situational requirements, to the ability to recognize the meaning of items, and to the ability to fake in line with the situational requirements has actually been demonstrated (Pauls & Crost, 2005). A further model expanded these aspects by suggesting that faking is likely to vary depending on the specific nature of the personality item (instead of considering the entire personality test), and including the factor "perceived ability to fake" which is more relevant to the motivation or intention to fake than the true ability to fake (Goffin & Boyd, 2009).

Two further theories, described by Hogan, Barrett, and Hogan (2007), are the self-report theory of faking and the impression management theory. The self-report theory describes the process of faking as inaccurate reports about the match between the content of an item and the content of memory, and assumes that a socialisation process involves training people to fake. The impression management theory argues that faking involves distorting the way one normally "communicates about oneself" (Hogan, Barrett, & Hogan, 2007) as people try to maximize acceptance and status and minimize rejection and the loss of status during social interaction. According to this theory the socialisation process involves training people in appropriate forms of self-presentation. Therefore it seems impossible to distinguish between faking and socialized behaviour if personality measures are administered that basically sample socialized adult behaviour and if (according to this theory) people know social norms rather than their real disposition. While the self-report theory predicts that the personality scores of people who are honest will be more consistent than scores of people who are dishonest, the impression management theory predicts that scores or people with good impression management skills will be more consistent than scores of people with poor impression management skills. While there is no evidence in research for the self-report theory (and it is inconsistent with research regarding how memory works as well as with modern theories about the nature of communication), there is support for the impression management theory (Hogan, Barrett, & Hogan, 2007).

Regarding the information processing, it is suggested that intentional response distortions rely on the comparison of the item content with an adopted schema (instead of a self-schema) like a schema of favourable impressions (Holden & Hibbs, 1995; Holden, Kroner, Fekken, & Popham, 1992). A systematic relationship between the presence of a schema and response time latencies is proposed. Thus, schema-congruent responses (e.g. socially desirable responses by adopting a fake good schema) are given faster than schema-incongruent responses (e.g. socially undesirable responses with regard to a fake good schema). But the resulting response latencies may also be influenced by item characteristics (item length, item ambiguity, item extremity, number of alternatives) and multiple person variables (reading speed, verbal ability, motor speed) (see also Holden, Fekken, & Cotton, 1991). Furthermore, it was shown that test-takers who were instructed to fake good on a personality questionnaire used job stereotypes (or schemas) however with negative aspects removed, as well as stereotypes about general aspects of personality (Mahar, Coburn, Griffin, Hemeter, Potappel, Turton, & Mulgrew, 2006; Mahar, Cologon, & Duck, 1995). The latter was assumed because training test-takers had no impact on the use of stereotypes in contrast to test-takers who were not trained.

Context effects and faking

Models of impression management show, that faking good can be defined as a context effect moderated by different interacting variables. We assume that faking good can also be moderated by other context effects (like fatigue or learning effects). With respect to the current experiment where the effects of an extensive test length (containing different contents) on intentional response distortion is investigated, context effects concerning item position, as well as learning and fatigue effects are supposed to occur and interact with the context effect of faking behaviour.

By investigating effects of different administration modes on intentional response distortion by instructing test-takers to fake good, it was revealed that test-takers were more likely to fake their answers at the beginning rather than at the end of a questionnaire (Seiwald, 2002). It was suggested that they might have forgotten the initial faking instructions over time. Another study revealed that certain questionnaire scales (Neuroticism and Conscientiousness) were shown to be less susceptible to socially desirable responses when items are randomly placed instead of grouped together (McFarland, Ryan, & Ellis, 2002).

According to the content of the measure, faking behaviour varies depending on the scale used (Griffin, Hesketh, & Grayson, 2004), and it seems as if specific dimensions (of the big five) are vulnerable to intentional response distortions like Neuroticism or Emotional Stability, Agreeableness, and Conscientiousness. Moreover it was shown that some of these scales correlate highly with socially desirable responding (Ferrando, 2008; Kurzt, Tarquini, & Iobst, 2008; McFarland & Ryan, 2000; Ones, Viswesvaran & Reiss, 1996; Rosse, Stecher, Miller, & Levin, 1998; Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). It was assumed that many of the items included in these dimensions are of socially desirable nature and therefore transparent to the test-takers. Conscientiousness has received the most attention in personnel selection literature and seems particularly susceptible to context effects (Griffin, Hesketh, & Grayson, 2004; Henry & Raju, 2006), but it also seems to be a strong predictor of job performance (Gill & Hodgkinson, 2007). As Conscientiousness, like Emotional Stability (or Neuroticism), is believed to be job-relevant (Ziegler & Bühner, 2009) it might be more important to test-takers to gain a beneficial score in these scales. Likewise, the dimension Order seems to be, maybe even more so, obviously job-related and was shown to be highly susceptible to impression management as well (Henry & Raju, 2006). This leads to the hypothesis that scales with obvious job-related content primes individuals to engage in impression management. It

was further assumed that traits or dimensions that have some positive as well as some negative aspects, like Extraversion (in contrast to dimensions that have almost negative associations, like Neuroticism), are more difficult to fake (Furnham, 1986). Faking did also appear to differ across facets of the Openness to Experience measure (Griffin, Hesketh, & Grayson, 2004), which was expected to be the least fakable measure of the Big Five (McFarland & Ryan, 2000). Scales that are perceived by applicants to be important to job performance are more likely to be faked than scales that seem to be less related to job performance (Zickar, Gibby, & Robie, 2004). It was shown that items with a context (e.g. work- or school related contents) specific wording (e.g. "at work" or "at school" tags) led to more positive responses than items with no context specific wording (without these tags); but items with context specific wording were also shown to have the highest criterion validity in a "behave-like-applicant" condition in contrast to a honest condition, than items without context specific wordings (Schmit, Ryan, Stierwalt, & Powel, 1995). It was therefore assumed that giving applicants a frame of reference might enhance the validity of a measure.

Some studies found increased reliabilities towards the end of a questionnaire (by investigating volunteers) which could be interpreted as a kind of learning effect because test taking induces self-awareness, or the respondents gain more knowledge about the test construct, or both (Hartig, Hölzel, & Moosbrugger, 2007; Knowles, 1988).

Aim of the current experiment

The aim of the current experiment is to provide a contribution to the research of intentional response distortion by investigating the influence of a context factor on this behaviour which has so far not been investigated: the item position with respect to the questionnaire length. Due to the question of whether different questionnaire contents or

dimensions are affected differently, the questionnaire used comprises items from three different questionnaires with items from the big five dimensions and job related items.

Hypotheses

According to the findings of learning effects towards the end of a questionnaire in non-applicant studies (Hartig, Hölzel, & Moosbrugger, 2007; Knowles, 1988) we assume that test-takers' ability to adjust their responses in a questionnaire to an adopted faking good schema (Holden & Hibbs, 1995; Holden, Kroner, Fekken, & Popham, 1992) might increase towards the end of a questionnaire if learning effects occur with respect to the content of the questionnaire scales. But in contrast to increasing item reliabilities in non-applicant studies, we assume that learning effects in selection situations lead to a decrease in scale reliabilities because the more applicants learn about the content of the questionnaire or get used to this kind of test, the higher the possibility that they can fit their responses to a certain stereotype might be (more equal or extreme responses within the sample towards the items). We therefore formulate the following hypotheses:

Hypothesis 1a: Intentional response distortion increases towards the end of a questionnaire.

Hypothesis 1b: Increasing response distortion towards the end of a questionnaire results in decreasing scale reliabilities as responses of the sample become more stereotypical (and therefore more equal) and less variable.

Referring to the findings of Seiwald (2002), who found that test-takers were more likely to fake their answers at the beginning rather than at the end of a questionnaire, we further assume that the adjustment of responses in questionnaires to an adopted faking good schema might require more concentration than referring to a self-schema. Therefore the ability to fake might decrease towards the end of an overlong questionnaire as fatigue effects might occur decreasing the required concentration or alertness. In contrast to

Hypothesis 1b, we assume that fatigue effects might result in increased scale reliabilities towards the end of a questionnaire because responses of the sample become less stereotypical (or equal) and more variable (as modal response categories are chosen more often). We therefore formulate a second hypothesis:

Hypothesis 2a: Intentional response distortion decreases towards the end of an overlong questionnaire.

Hypothesis 2b: Decreasing response distortion towards the end of a questionnaire results in increasing scale reliabilities as responses of the sample become more variable and less stereotypical.

Based on the fact that socially desirable responding seems to affect particular dimensions like Neuroticism, Emotional Stability, Agreeableness, and Conscientiousness (McFarland & Ryan, 2000; Ones, Viswesvaran & Reiss, 1996; Rosse, Stecher, Miller, & Levin, 1998), and the opinion that personality tests can provide substantial criterion validity when used in personnel selection if the measured traits are matched to the nature of the job (Goffin & Boyd, 2009), we formulate a third hypothesis:

Hypothesis 3: If the assumed effects described in Hypothesis 1 and Hypothesis 2 occur, they occur rather in a questionnaire that measures the big five dimensions than in a questionnaire with general job related contents.

Method

An experiment within a selection procedure is presented to investigate the effects of item position on faking personality questionnaires with regard to the questionnaire length. The sample consists of soldiers in the Austrian Federal Armed Forces who applied for pilot training and went through a selection procedure.

Sample

All applicants were soldiers in the Austrian Federal Armed Forces who had applied for pilot training. A total of 84 applicants, 1 female and 83 male, between the age of 18 and 23 filled out a paper-pencil questionnaire consisting of 516 items after an assessment of approximately 8 to 9 hours to test their cognitive abilities and personality traits. Their first language is German, their education levels varies from a compulsory education of 9 years (n = 11) and apprenticeship (n = 36) to general qualification for university entrance (n = 37).

Preliminary stages

First of all, the requirements profile of the Department of Human Resources of the Austrian Federal Armed Forces was considered in order to ascertain the requirements the pilots were supposed to fulfil. The Department of Human Resources established the following personality traits to be of particular importance with regard to the requirements for pilots: achievement motivation, willingness to learn, mobility, sense of responsibility, decision-making ability, openness to contact, extraversion, ability to communicate ideas and feelings in a socially acceptable way, dominance, empathy as the ability to understand feelings and experiences of other people, aggression as the disposition to defend oneself against offences and unfairness (aggression in the sense of the absence of self-control is not desired), emotional stability and stress management (cognitive and behavioural ones).

Based on these requirements the German edition of the Personality Research Form (PRF; Stumpf, Angleitner, Wieck, Jackson, & Beloch-Till, 1984), a questionnaire that was already included in the assessment of pilots by the Department of Human Resources of the Austrian Federal Armed Forces, was enhanced with items from three additional personality questionnaires: the German edition of the Business-focused Inventory of Personality (BIP; Hossiep & Paschen, 2003), the German edition of the NEO Five Factor Inventory (NEO FFI; Borkenau & Ostendorf, 1993), and the German edition of the NEO Personality

Inventory Revised (NEO PI-R; Ostendorf & Angleitner, 2004). Hence, a personality questionnaire with 516 items was compiled for the sample of Study 1.

As the effects of questionnaire length or item position on faking tendencies (in the sense of social desirability) should be investigated, it needed to be decided, which scores with respect to the different questionnaire scales might be indicators of faking tendencies. According to the Department of Human Resources of the Austrian Federal Armed Forces the following scores are requested and may therefore be interpreted as indicators of faking good tendencies: 1) high scores of achievement motivation, willingness to learn, mobility, sense of responsibility, decision-making ability, openness to contact, extraversion, ability to communicate ideas and feelings in a socially acceptable way, empathy as the ability to understand feelings and experiences of other people, emotional stability, and stress management; 2) average scores of dominance, and aggression as the disposition to defend oneself against offences and unfairness; 3) low scores of aggression in the sense of the absence of self-control is not desired.

Measures

According to the requirements profile items from the following personality questionnaires were used to put together an appropriate personality questionnaire.

Business-focused Inventory of Personality (BIP) – German edition. The BIP (Hossiep & Paschen, 2003) is a work-based personality questionnaire that combines an assessment of work style and motivation. Fourteen scales arranged into four conceptual domains are measured with 210 items by responding to a rating scale with 6 categories (from “agree completely” to “disagree totally”): 1) “Occupational Orientation” assesses work specific motivation with three scales: Achievement Motivation, Power Motivation, Leadership Motivation; 2) “Occupational Behaviour” assesses the typical approach to work with three scales: Conscientiousness, Flexibility, Action Orientation; 3) “Social

Competencies" describes the style of interacting with other people and contains five scales: Social Sensitivity, Openness to Contact, Sociability, Team Orientation, Assertiveness; 4) "Psychological Constitution" assesses how the demands made by a range of tasks at work impact on a person's resilience and experience of emotional pressure, and contains three scales: Emotional Stability, Working under Pressure, Self-Confidence. Four additional scales can be generated from some of the above mentioned items: "Additional Index Mobility", "Additional Index Orientation towards Leisure-Time", "Additional Index Control Experience", and "Additional Index Competitive Orientation". The cronbachs alpha and split-half reliabilities of the BIP scales presented in the test-manual range from .72 to .91.

NEO Personality Inventory Revised (NEO PI-R) – German edition. The NEO PI-R (Ostendorf & Angleitner, 2004) is a personality questionnaire that measures the five major domains of personality with 240 items by responding to a rating scale with five categories (from "strong agreement" to "strong disagreement", and the middle category "neutral"). Each domain (scale) is divided into six facets (subscales): 1) "Neuroticism" identifies individuals who are prone to psychological distress: Anxiety, Angry Hostility, Depression, Self Consciousness, Impulsiveness, Vulnerability; 2) "Extraversion" measures the quantity and intensity of energy directed outwards into the social world: Warmth, Gregariousness, Assertiveness, Activity, Excitement seeking, Positive Emotion; 3) "Openness to Experience" measures if someone actively seeks and appreciate experiences for their own sake: Fantasy, Aesthetics, Feelings, Actions, Ideas, Values; 4) "Agreeableness" measures the kinds of interactions an individual prefers ranging from compassion to tough mindedness: Trust, Straightforwardness, Altruism, Compliance, Modesty, Tender mindedness; 5) "Conscientiousness" measures the degree of organization, persistence, control and motivation in goal directed behaviour: Competence, Order, dutifulness, Achievement striving, Self Discipline, Deliberation. According to the above mentioned

requirements profile only the subscales Angry Hostility (subscale of Neuroticism) and Compliance (subscale of Agreeableness) were chosen in addition to the NEO FFI described below. The cronbachs alpha reliabilities of these two subscales presented in the test-manual range from .69 to .79.

NEO Five Factor Inventory (NEO FFI) – German edition. The NEO FFI (Borkenau & Ostendorf, 1993) is a short version of the NEO PI-R that provides a brief measure of the five domains Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness. These five major domains of personality are measured with 60 items by responding to a rating scale with five categories (from "strong agreement" to "strong disagreement", and the middle category "neutral"). The cronbachs alpha reliabilities of the NEO FFI scales presented in the test-manual range from .67 to .85.

Personality Research Form (PRF) – German edition. The PRF (Stumpf, Angleitner, Wieck, Jackson, & Beloch-Till, 1984) is a personality questionnaire based on the personality theory of Murray (1938) that measures a set of traits important for psychological research as well as psychological assessment. 234 items with a dichotomous response format ("right" and "wrong") measure fifteen scales: Achievement, Affiliation, Aggression, Dominance, Endurance, Exhibition, Harm avoidance, Impulsivity, Nurturance, Order, Play, Social Recognition, Succorance, Understanding. The cronbachs alpha reliabilities of the PRF scales presented in the test-manual range from .69 to .87.

The final personality questionnaire for Study 1 comprises 39 scales and a total of 516 items (210 items of the BIP, 60 items of the NEO FFI, 12 items of the NEO PI-R, and 234 items of the PRF) presented with a rating scale with six categories. This response format was chosen to avoid response behaviour with a tendency to always choose the middle category, and to allow a more detailed self-report than for example a dichotomous response format does in order to avoid reactance (Karner, 2002).

Procedure

First, all applicants attended a psychological assessment by the Department of Human Resources of the Austrian Federal Armed Forces where their cognitive abilities and personality traits were tested. After an assessment of approximately 8 to 9 hours (including cognitive ability and achievement tests), they filled out the paper-pencil questionnaire consisting of a total of 516 items (of the BIP, NEO PI-R, NEO FFI, and PRF) by responding to a rating scale with six categories. None of the applicants received information about the requirements profile. The applicants were randomly assigned to the two experimental groups.

Design

To test whether the questionnaire length or item position has any effects, 84 participants completed a paper-pencil questionnaire with a total of 516 items composed of the items fro the BIP, NEO FFI, NEO PI-R and PRF in a unifactorial multivariate design. The questionnaire was divided into three parts that were combined into two different item orders to investigate the item positions: "Part 1 – Part 2 – Part 3" (Item Position B-N) and "Part 3 – Part 2 – Part 1" (Item Position N-B). Part 1 (129 items from the BIP – in the following labelled as "BIP1") and Part 3 (60 items from the NEO PI-R, and 12 items of the NEO FFI) were varied to change the item position, and included 18 scales that were the most important ones with regard to the requirements profile. Part 2 (81 items from the BIP – in the following labelled as "BIP2" – and 234 items from the PRF) included the remaining 21 scales and was to be the constant middle part that was not varied. The item order within each of the three parts was not varied but held constant. Thereby, the original item order of the BIP, PRF, and NEO FFI was kept, while the items of the two NEO PI-R subscales were interspersed between the NEO FFI items with constant distance. The participants were randomly assigned to the two experimental groups: one group completed the questionnaire with the item position "B-N", and one group completed the questionnaire

with the item position "N-B". The experimental design (the distribution of the scales and items to the three parts) is given in Figure 1.

Insert Figure 1 about here

Results

To investigate Hypotheses 1a and Hypothesis 2a a multivariate analysis of variance (MANOVA) was conducted to compare means of the two experimental groups in regard to the main factor Item Position. To calculate the sample sizes needed to fulfil the a-priori given precision requirements (type-I, type-II-risk, and a relevant effect size) we used the program CADEMO (http://www.biomath.de). As a matter of fact we had to calculate the sample sizes according to ANOVA. With $\alpha = .05$ and $\beta = .20$ an ANOVA is able to detect a mean difference of $\delta \geq 2/3\ \sigma$ (the standard deviation of each scale) by testing 37x2 = 74 subjects. With 42x2 = 84 we had achieved/obtained/realised adequate sample sizes. Additional to a MANOVA we calculated the cronbachs alpha reliabilities of all scales of the BIP, NEO FFI, and NEO PI-R for each experimental group to investigate Hypothesis 1b and Hypothesis 2b.

Results of the MANOVA

The means and standard deviations of all subtests in each experimental condition are given in Table 1. Box's M-Test for testing the homogeneity of the variance-covariance matrix was significant ($p = .048$). To find out if this significance is due to particular dependent variables (questionnaire scales) because of heterogeneous variances, the Levene's test was calculated for each scale. One scale of the BIP1 (Achievement Motivation) was disclosed to be significant in the Levene's test ($p = .033$) and was

therefore excluded. Because the Box's M-Test on the remaining scales was still significant ($p = .018$), we compared Pearson correlation coefficients of all dependent variables in order to detect any differences in scale covariances between the experimental groups. The greatest differences between the two experimental groups were shown by the correlation of the BIP scale Emotional Stability with the BIP scale Conscientiousness with correlation coefficients .310 and -.139, and by the correlation of the BIP scale Emotional Stability with the BIP scale Openness to Contact with correlation coefficients .604 and .093. After deleting the scale Emotional Stability Box's M-Test was still significant ($p = .010$). Then, the NEO PI-R scale Angry Hostility was deleted next as it showed the next greatest difference between the two experimental groups by correlating with the BIP scale Working under Pressure with correlation coefficients -.789 and -.370. After deleting this scale Box's M-Test proved to be non significant ($p = .079$). That is, the resulting *F*-values of the multivariate analysis of variance can be fairly interpreted.

The MANOVA showed a significant effect of Item Position ($p = .014$; $F = 2.218$; Hypothesis $df = 15$; Error $df = 68$; $\eta^2 = .329$). The separate invariate analyses of the each single scale showed that only one scale of the NEO FFI (Conscientiousness) revealed significantly different means between the two experimental groups ($p = .031$, $F = 4.838$, $\eta^2 = .056$). See in Table 1 the respective means. In order to further investigate the factor Item Position on the BIP scale Achievement Motivation and the NEO PI-R scale Angry Hostility, which had to be deleted from the MANOVA, two-sample *t*-tests for unequal variances (Welch tests) were applied: no significant effect occurred ($p = .489$, $p = .709$)

Insert Table 1 about here

Cronbachs alpha reliability estimates

The cronbachs alpha reliabilities for all experimental groups are given in Table 2. The reliabilities of almost all questionnaire scales show a common trend: scales of the BIP, NEO FFI, and NEO PI-R which were applied towards the end of the questionnaire showed higher reliabilities than when they were applied at the beginning of the questionnaire. Except for the NEO FFI scale Agreeableness and the NEO PI-R scale Angry Hostility, whose reliabilities tended to be higher at the beginning than at the end of the questionnaire.

Insert Table 2 about here

Interpretation

The results of MANOVA provide evidence so that Hypothesis 2a is accepted, but Hypothesis 1a is rejected. Item position (questionnaire length) affected subjects' responses according to one scale of the NEO FFI. If items of the NEO FFI were presented later in the questionnaire, subjects described a significantly lower conscientiousness than subjects who answered the same items at the beginning of the questionnaire (cf. Table 1). If the cronbach alpha reliability estimates are considered then the common trend that scale reliabilities tended to be higher at the end than at the beginning of the questionnaire supports Hypothesis 2b, excluding the NEO FFI scale Agreeableness and the NEO PI-R scale Angry Hostility, whose reliabilities tended to be higher at the beginning than at the end of the questionnaire, thereby supporting Hypothesis 1b. Overall, the results show that a fatigue effect might have occurred, decreasing the concentration or alertness, thereby also decreasing the applicants' ability to adjust their responses in questionnaires to an adopted faking good schema.

Discussion

An experiment with a multivariate two-way design was conducted in order to investigate the influence of item position with regard to a questionnaire's length on intentional response distortion. A questionnaire with 516 items was administered to applicants who had applied for pilot training in the Austrian Federal Armed Forces. The questionnaire was composed of items from the BIP, PRF, NEO FFI, as well as NEO PI-R, and was administered beginning either with items from the BIP and ending with items from the NEO FFI and NEO PI-R, or with the reversed item order (the items of the PRF were always positioned in the middle with regard to both item orders). The significant main effect of MANOVA provides evidence for Hypothesis 2a, but Hypothesis 1a is rejected. Item position (questionnaire length) affected subjects' responses to the scale Conscientiousness in the NEO FFI. Subjects describe a significantly lower conscientiousness at the end of an overlong questionnaire than subjects who receive the same items at the beginning of a questionnaire. The findings resemble those of Seiwald (2002) who found that test-takers were more likely to fake their answers at the beginning rather than at the end of a questionnaire. One possible explanation for this finding might be a decreased alertness or concentration towards the end of the questionnaire as a result of fatigue effects, making it more difficult for test-takers to compare the item contents with a faking schema.

As the effects occurred only with respect to the NEO FFI scale Conscientiousness, while none of the BIP scales was affected by the item position, evidence for Hypothesis 3 is provided. The big five dimensions seem to be more vulnerable to intentional response distortions than general job related dimensions. With that said, the findings are in line with the common literature that revealed that faking affects particular dimensions like Neuroticism, Emotional Stability, Agreeableness, and Conscientiousness (McFarland & Ryan, 2000; Ones, Viswesvaran & Reiss, 1996; Rosse, Stecher, Miller, & Levin, 1998).

Of course, the fact that only the dimension Conscientiousness was affected and not any other big five dimension might also lead to the conclusion that the findings are only coincidental. But, we assume that the items of the dimension Conscientiousness might be even more transparent with regard to the measured content than items from other big five dimensions, making it more easy to choose socially desirable responses at the beginning of a questionnaire when concentration and alertness are still given. Additionally, the applicants had no information about the requirements profile and therefore no information as to what was desirable with regard to the job/training they had applied for. As test-takers who have knowledge about desirable behaviours with respect to the aspired job would be better able to fake their responses (Goffin & Boyd, 2009; Levashina & Campion, 2006; Snell, Sydell, and Lueke, 1999), it might have been more difficult for the applicants in the current experiment to figure out how to respond in a desirable way with respect to other scales except the scale Conscientiousness. Because the scale Conscientiousness might have been more susceptible to intentional response distortion at the beginning of the questionnaire for the mentioned reasons, we assume that this scale was therefore also more prone to the fatigue effect at the end of the questionnaire, resulting in significant mean differences.

Though, only the scale Conscientiousness revealed significant mean differences, the cronbach alpha reliability estimates for almost all questionnaire scales except two, provide support for Hypothesis 2b which partially supports Hypothesis 2a. As scale reliabilities tend to be higher at the end rather than at the beginning of the questionnaire it can be assumed that socially desirable response distortion decreases towards the end of the questionnaire as a result of decreased concentration or alertness (fatigue effect).

The fact that the NEO FFI scale Agreeableness and the NEO PI-R scale Angry Hostility showed higher reliabilities at the beginning rather than at the end of the questionnaire could be interpreted as learning effects on the one hand, thereby showing

that it might have been easier for applicants to fit their responses to a faking good schema (Hypothesis 1b). However, on the other hand, these findings could be interpreted as a kind of frustration effect due to the overlong questionnaire which was applied additionally after an 8-9 hour long assessment (of course with breaks in between). When the means of the two experimental groups are compared it is evident that the mean of the scale Agreeableness was lower and the mean of the scale Angry Hostility was higher when these scales were applied at the end of the questionnaire than when they were applied at the beginning. This would tend to support the hypothesis that some kind of frustration occurred at the end of the questionnaire, rather as an expression of the actual emotional state than as an expression of faking tendencies.

Limitations and implications for further research

One limitation of the current study is that all participants were men, who had all applied for the same (pilot) training and who all came from the same institution (Austrian Federal Armed Forces). This makes the sample a very special one, and the findings might not apply to other groups, such as women, or other occupational groups. Future research should investigate if the same effects can be found in other samples. As faking is assumed to be a variable of individual differences (Mueller-Hanson, Heggestad, & Thornton, 2006; Viswesvaran, & Ones, 1999), we cannot maintain that all applicants distorted their responses to the same extend. Furthermore, it has to be considered that the applicants received no information about the requirements profile which might have decreased their ability to fake (Goffin & Boyd, 2009; Levashina & Campion, 2006; Snell, Sydell, and Lueke, 1999). Our hypothesis of fatigue effects decreasing the ability to fake by making it more difficult to fit responses to a faking schema should be ascertained in further research by measuring response latencies per item with a computer based questionnaire. In this way,

it could be investigated whether schema congruent responses are given faster at the beginning of a questionnaire rather than at the end.

Acknowledgments

We are grateful to the Department of Human Resources of the Austrian Federal Armed Forces, in particular Michael Mikas (Head of the Ambulance for Aviation & Traffic Psychology), Christian Czihak (Head of the Department for Aviation & Traffic Psychology), and Christian Langer (Head of the Department of Human Resources), for granting permission to test applicants within the local selection procedure.

References

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren Inventar (NEO-FFI)* [*NEO Five Factor Inventory (NEO FFI)*]. Manual, Göttingen: Hogrefe.

Bradley, K. M., & Hauenstein, N. M. A. (2006). The moderating effects of sample types as evidence of the effects of faking on personality scale correlations and factor structure. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*]*, 48*, 313-335.

Dilchert, S., Ones, D. S., Viswesvaran, C., & Deller, J. (2006). Response distortion in personality measurement: born to deceive, yet capable of providing valid self-assessments? *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*]*, 48*, 209-225.

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*, 155-166).

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.

Ferrando, P. J. (2008). The impact of social desirability bias on the EPQ-R item scores: An item response theory analysis. *Personality and Individual Differences, 44*, 1784-1794.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences, 7*, 385-400.

Gill, C. M., & Hodgkinson, G. P. (2007). Development and validation of the Five-Factor Model Questionnaire (FFMQ): An adjectival-based personality inventory for use in occupational settings. *Personnel Psychology, 60*, 731-766.

Goffin, R. D., & Boyd, A. C. (2009). Faking and personality assessment in personnel selection: Advancing models of faking. *Canadian Psychology, 50*, 151-160.

Griffin, B., Hesketh, B., & Grayson, D. (2004). Applicants faking good: evidence of item bias in the NEO PI-R. *Personality and Individual Differences, 36*, 1545-1558.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behaviour. *Personnel Review, 36*, 341-355.

Hartig, J., Hölzel, B., & Moosbrugger, H. (2007). A Confirmatory Analysis of Item Reliability Trends (CAIRT): Differentiating true score and error variance in the analysis of item context effects. *Multivariate Behavioral Research, 42,* 157-183.

Henry, M. S., & Raju, N. S. (2006). The effects of traited and situational impression management on a personality test: an empirical analysis. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*], *48*, 247-267.

Hogan, J., Barrett, P. & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92*, 1270-1285.

Holden, R., & Hibbs, N. (1995). Increment validity of response latencies for detecting fakers on a personality test. *Journal of Research in Personality, 29,* 362-372.

Holden, R. R., Fekken, G. C., & Cotton, D. H. G. (1991). Assessing psychopathology using structured test-item response latencies. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3,* 111-118.

Holden, R. R., Kroner, D. G., Fekken, G. C., & Popham, S. M. (1992). A model of personality test item response dissimulation. *Journal of Personality and Social Psychology, 63,* 272-279.

Hossiep, R., & Paschen, M. (2003). *Das Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP)* [*Business-focused Inventory of Personality (BIP)*]. Manual, 2., vollständige überarb. Auflage, Göttingen: Hogrefe.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581-595.

Karner, T, (2002). The volunteer effect of answering personality questionnaires. *Psychologische Beiträge* [latterly: *Psychological Test and Assessment Modeling*], *44,* 42-49.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*, 312-320.

Kurtz, J. E., Tarquini, S. J., & Iobst, E. A. (2008). Socially desirable responding in personality assessment: Still more substance than style. *Personality and Individual Differences, 45,* 22-27.

Levashina, J., & Campion, M. A. (2006). A model of faking likelihood in the employment interview. *International Journal of Selection and Assessment, 14*, 299-316.

Mahar, D., Cologon, J., & Duck, J. (1995). Response strategies when faking personality questionnaires in a vocational selection setting. *Personality and Individual Differences, 18*, 605-609.

Mahar, D., Coburn, B., Griffin, N., Hemeter, F., Potappel, C., Turton, M, & Mulgrew, K. (2006). Stereotyping as a response strategy when faking personality questionnaires. *Personality and Individual Differences, 40,* 1375-1386.

Marcus, B. (2003a). Das Wunder sozialer Erwünschtheit in der Personalauswahl. [The wonder of social desirability in personnel selection settings] *Zeitschrift für Personalpsychologie, 2,* 129-132.

Marcus, B. (2003b). Persönlichkeitstests in der Personalauswahl: Sind „sozial erwünschte" Antworten wirklich nicht wünschenswert? [Personality testing in personnel selection: Is „socially desirable" responding really undesirable?]. *Zeitschrift für Personalpsychologie, 2,* 138-148.

Marcus, B. (2006). Relationships between faking, validity, and decision criteria in personnel selection. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*], *48*, 226-246.

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*, 812-821.

McFarland, L. A., & Ryan, A. M. (2006). Toward an integrated model of applicant faking behavior. *Journal of Applied Social Psychology, 36*, 979-1016.

McFarland, L. A., Ryan, A. M., & Ellis, A. (2002). Item placement on a personality measure: Effects on faking behaviour and test measurement properties. *Journal of Personality Assessment, 78,* 348-369.

Mueller-Hanson, M., Heggestad, E. D., & Thornton, G. C. (2003). Faking and selection: Considering the use of Personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88,* 348-355.

Mueller-Hanson, R. A., Heggestad, E. D., & Thornton, G. C. III (2006). Individual differences in impression management: An exploratory of the psychological process underlying faking. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*], *48*, 288-312.

Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test or the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology, 65,* 131-149.

Morgeson, F. P., Campion, M. A., Diboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683-729.

Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995-1027.

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.

Ones, D.S., Viswesvaran, C. &, Schmidt, F.L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78,* 679-703.

Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae revidierte Fassung (NEO-PI-R)* [*NEO Personality Inventory Revised according to Costa and Mc Crae (NEO PI-R)*]. Manual, Göttingen: Hogrefe.

Pauls, C. A., & Crost, N. W. (2005). Cognitive ability and self-reported efficacy of self-presentation predict faking on personality measures. *Journal of Individual Differences, 26*, 194-206.

Robie, C., Brown, D. J., & Beaty, J. C. (2007). Do people fake on personality inventories? A verbal protocol analysis. *Journal of Business and Psychology, 21*, 489-509.

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.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Ressource Management Review, 16*, 155-180.

Schmit, M. J., Ryan, A. M., Stierwalt, S. L., & Powell, A. B. (1995). Frame-of-reference effects on personality scale scores and criterion-related validity. *Journal of Applied Psychology, 80,* 607-620

Seiwald, B. B. (2002). Replicability and generalizability of Kubinger's results: Some more studies on faking personality inventories. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*], 44, 17-23.

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.

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review, 9,* 219-242.

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.

Stumpf, H., Angleitner, A., Wieck, T., Jackson, D. N., & Beloch-till, H. (1984). *Deutsche Personality Research Form (PRF)* [*German Version of the Personality Research Form (PRF)*]. Manual, Göttingen: Hogrefe.

Tsaousis, I., & Nikolaou, I. E. (2001). The stability of the five-factor model of personality in personnel selection and assessment in Greece. *International Journal of Selection and Assessment, 9,* 290-301.

Viswesvaran, C., & Ones, D.S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197-210.

Winkelspecht, C., Lewis, P., & Thomas, A. (2006). Potential effects of faking on the NEO-PI-R: Willingness and ability to fake changes who gets hired in simulated selection decisions. *Journal of Business and Psychology, 21*, 243-259.

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.

Ziegler, M., & Bühner, M. (2009). Modeling socially desirable responding and its effects. *Educational and Psychological Measurement, 69*, 548-565.

Table 1:
Means and standard deviations for all dependent variables (scales) in the experimental condition Item Position (2 experimental groups)

Dependent Variable (Scale)	Item Position B-N (BIP1 – NEO) (n=42)		Item Position N-B (NEO – BIP1) (n=42)	
	Mean	Standard Deviation	Mean	Standard Deviation
Achievement Motivation – BIP1	64.857	6.606	63.679	8.775
Conscientiousness – BIP1	63.869	8.651	64.286	10.163
Action Orientation – BIP1	65.881	8.502	62.845	9.092
Social Sensitivity – BIP1	53.441	6.612	52.381	8.048
Sociability – BIP1	59.524	8.170	59.714	9.105
Openness to Contact – BIP1	73.071	8.744	68.988	11.354
Assertiveness – BIP1	51.321	6.030	50.357	7.133
Emotional Stability – BIP1	68.452	8.577	66.798	9.479
Working under Pressure – BIP1	64.821	7.465	61.500	9.719
Additional Index Mobility – BIP1	10.000	2.024	9.595	2.480
Additional Index Orientation towards Leisure-Time – BIP1	31.762	4.160	29.786	5.215
Neuroticism – NEO FFI	25.214	7.063	26.131	7.900
Extraversion – NEO FFI	55.131	6.990	55.369	6.402
Openness to Experience – NEO FFI	46.024	7.706	46.952	7.520
Agreeableness – NEO FFI	52.333	6.506	53.655	6.879
Conscientiousness – NEO FFI	61.738	6.420	64.619	5.552
Angry Hostility – NEO PI-R	20.060	4.575	19.643	5.578
Compliance – NEO PI-R	28.941	5.342	30.583	4.748

Table 2:
Cronbachs alpha reliability of the scales from the BIP, NEO FFI, and NEO PI-R for each experimental group

Dependent Variable (Scale)	Item Position B-N (BIP1 – NEO)	Item Position N-B (NEO – BIP1)
Achievement Motivation – BIP1	.638	.802
Conscientiousness – BIP1	.776	.879
Action Orientation – BIP1	.823	.853
Social Sensitivity – BIP1	.677	.837
Sociability – BIP1	.667	.774
Openness to Contact – BIP1	.736	.877
Assertiveness – BIP1	.629	.740
Emotional Stability – BIP1	.667	.793
Working under Pressure – BIP1		
Additional Index Mobility – BIP1	.687	.910
Additional Index Orientation towards Leisure-Time – BIP1	.443	.633
	Item Position N-B (NEO – BIP1)	Item Position B-N (BIP1 – NEO)
Neuroticism – NEO FFI	.768	.799
Extraversion – NEO FFI	.649	.795
Openness to Experience – NEO FFI	.653	.719
Agreeableness – NEO FFI	.726	.711
Conscientiousness – NEO FFI	.793	.849
Angry Hostility – NEO PI-R	.645	.510
Compliance – NEO PI-R	.415	.625

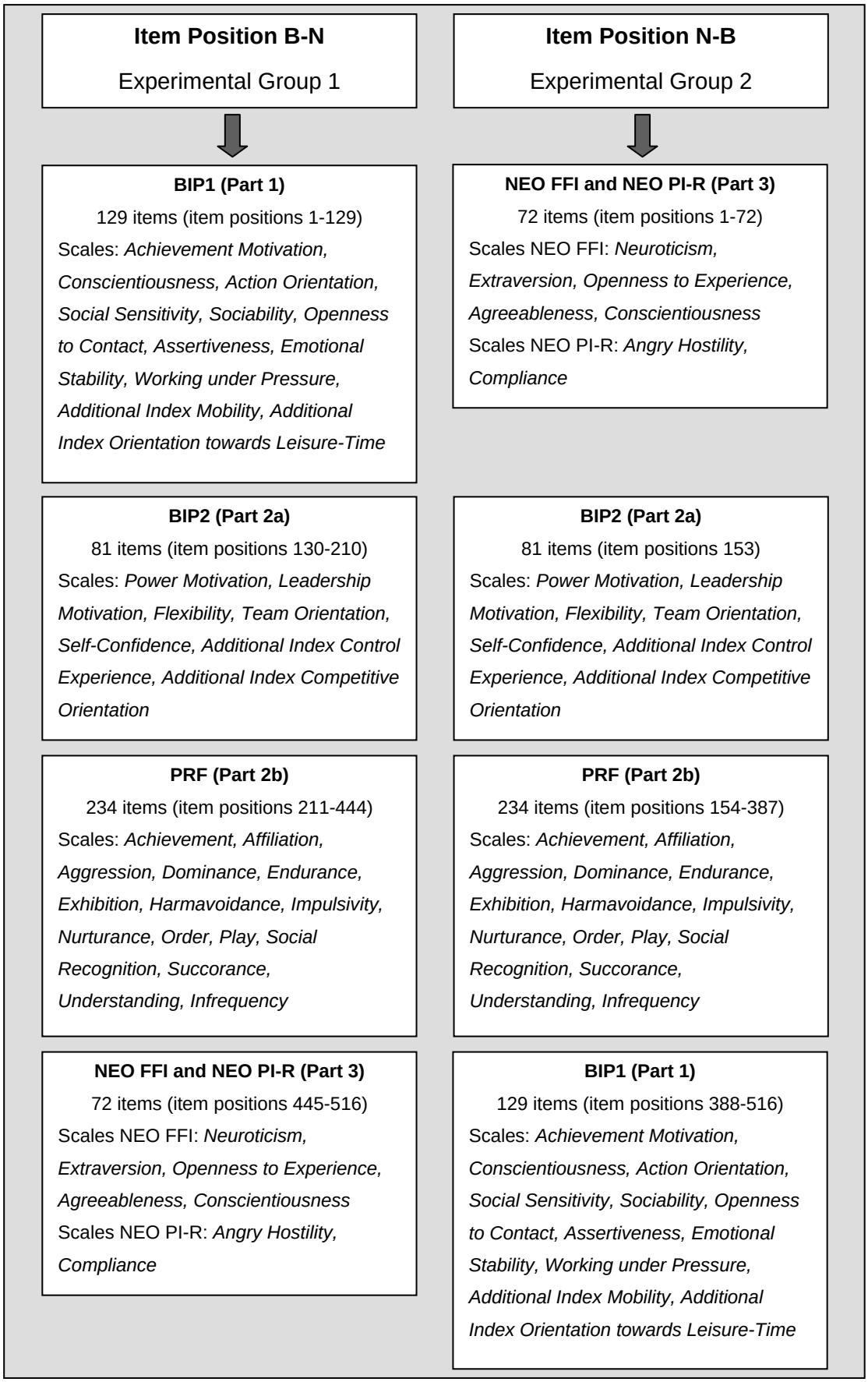


Figure 1: Experimental Design

5.3. Paper 3

Khorramdel, L., & Frebort, M. (accepted for publication). Context effects on test performance: What about test order? *European Journal of Psychological Assessment*.

Context effects on test performance: What about test order?

Lale Khorramdel and Martina Frebort

University of Vienna

Lale Khorramdel and Martina Frebort, Center of Testing and Consulting, Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna.

Correspondence concerning this article may be addressed to Lale Khorramdel, University of Vienna, Liebiggasse 5, 1010 Vienna, Austria, Email: lale.khorramdel@univie.ac.at, or Martina Frebort, Email: martina.frebort@univie.ac.at

Abstract

The effects of varied test order on test performance within a computer test battery were investigated. An experiment was performed to determine whether completing objective personality tests sensu R. B. Cattell affects test performance in subsequent cognitive ability tests and vice versa. The sample consisted of managers of an industrial corporation (an automotive supplier) in "higher management positions" (business managers, department chiefs, and team leaders) who attended an investigation of their professional potential which resembles a real selection situation. It was hypothesized that carry-over and priming effects, as well as fatigue and learning effects might occur. Results of a MANOVA showed a main effect of test order on objective personality tests, since "frustration tolerance" decreased and "decisiveness" increased when objective personality tests were presented subsequent to cognitive ability tests, while cognitive ability tests were not affected by prior objective personality tests.

Key words: context effects, test order, frustration tolerance, objective personality tests, cognitive ability tests, selection, assessment

Introduction

Context effects, especially concerning item order or item positions, are mainly investigated and reported in research involving interviews and questionnaires (Hartig, Hölzel, & Moosbrugger, 2007; Knowles, 1988; Knowles et al., 1992; Rost, & Hoberg, 1996). Fewer studies are addressed to task order of cognitive ability and achievement tests (Leary, & Dorans, 1985; Perlini, Lind, & Zumbo, 1998), and only a handful of studies deal with test order within test batteries (Baldinger, 2006; Földényi, Tagwerker-Neuenschwander, Giovanoli, Schallberger, & Steinhausen, 1999; Eiselt, 1991). Models of information processing are sometimes used to explain order effects. Such effects occur as different orders may influence a four-stage process (interpretation, retrieval of information, rendering a judgement, selection of a response) underlying response behaviour (Tourangeau, & Rasinski, 1988). This influence may not be limited to item order but may also extend to test order. In the following, order effects in personality questionnaires, as well as cognitive ability tests are presented to help determine what kinds of context effects due to test order might occur within a test battery. Furthermore, relevant studies focusing on test order are presented. But first, a short overview of different context effects is provided.

Context effects

An overview of different context effects is given by Smith (1992). In our opinion, mainly carry-over effects, priming effects, fatigue effects, and learning effects are the focus of interest in regard to different test orders, as they involve the transfer of prior content, meaning, or behaviour and influence subsequent reactions. While carry-over effects, priming effects, and learning effects are related to the contents of prior items or a prior experimental condition, position effects like fatigue effects or the increase of openness

during an interview (or the decrease of anxiety during a test situation) are not. Schwarz and Wyer (1983) assume that anchoring effects occur because responses are not always determined by mapping subjective beliefs directly on a response scale but by evaluating stimuli in relation to one another. This could be broadened to carry-over effects and priming effects as well. The occurrence of priming effects seems to be enhanced by subjects' exposure to the priming objects through a direct, behavioural experience (direct priming), in contrast to indirect priming that does not include behavioural experience (Fazio, Powell, & Herr, 1983). Furthermore, different kinds of priming induced by different kinds of tasks may enhance different problem-solving strategies in subsequent tasks. La Rue and Olejnik (1980) showed that formal operational priming through previously presented verbal seriation tasks enhanced formal operational strategies, thereby facilitating moral reasoning. In contrast, working on previously presented addition and subtraction tasks (concrete operational priming) led to lower scores in moral reasoning. Such effects due to the content of prior tasks might be expected in test batteries due to the content of tests as well.

Context effects in personality and cognitive ability measures

Although there are a lot of differences between personality and cognitive ability measures - for example the aim of the measure (attitudes versus abilities) or the used techniques (questioning versus testing) – the same types of context effects seem to occur: foremost learning or fatigue effects. While learning effects in personality questionnaires are revealed by increasing item reliabilities towards the end of a questionnaire (Knowles, 1988), learning effects in cognitive ability tests tend to result in higher test scores or lower item difficulties towards the end of a test (Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007; Kubinger, 2009a). Fatigue effects due to later item positions in personality questionnaires result in more blanks towards the end of the questionnaire (Kraut, Wolfson,

& Rothenberg, 1975) while they lead to slowed down reactions or increased errors in cognitive ability tests (Földényi, Tagwerker-Neuenschwander, Giovanoli, Schallberger, & Steinhausen, 1999); both response behaviours result in lower scores. A comparison of studies investigating effects of different item orders within personality questionnaires (mainly randomized item sequences versus homogeneous item blocks) shows that those studies which found effects often used questionnaires or scales that were not mainly focussed on job or achievement related contents, such as locus-of-control scales, anxiety inventories, scales that measure extraversion, neuroticism and masculinity or symptom checklists (Franke, 1997; Knowles, 1988; Knowles, & Byers, 1996; Steinberg, 1994; Ortner, 2008). In contrast, questionnaires that measure academic self concept of achievement and ability (Rost, & Hoberg, 1996; Sparfeldt, Schilling, Rost, & Thiel, 2006), or job-related contents (Baehr, 1953; Schriesheim, 1981; Schriesheim, Kopelman, & Solomon, 1989) show weaker effects due to item order or no effect at all.

Effects of test order

The most interesting studies with respect to the current experiments are, of course, studies that focus on varied test sequences. Comparing a randomized with a blocked (items of the same scale are stringed together and presented in blocks) item order within a Rasch model fitting Big Five personality questionnaire and within an achievement motivation questionnaire revealed no effects, nor did the test order show any effects by varying these two questionnaires (Eisenhauer, 2008). Varying the sequence of cognitive ability tests (attention tests) and personality questionnaires no effects of personality questionnaires on cognitive ability tests were found when administered first, but two questionnaire scales (spontaneous aggressiveness, emotional lability) showed higher means when cognitive ability tests were administered before the questionnaires (Hambros, 2002). Arranging

personality questionnaires after objective personality tests (experiment-based assessments of behaviour) did not affect the questionnaire scales, but the test score "decisiveness" of the objective personality test *Work Styles* ("Arbeitshaltungen"; Kubinger, & Ebenhöh, 1996; adapted 2007) was significantly higher when administered after a personality questionnaire that measured job-related attitudes (Baldinger, 2006). The investigation of different orders of cognitive ability tests within a computer based test battery revealed the same response behaviour in each of the different test orders: increased working speed as well as a decreased number of right answers towards the end of the test battery (Eiselt, 1991). These findings were ascribed more to self-reported motivation than to self-reported subjective fatigue, similar to findings of Ackerman and Kanfer (2009). Both fatigue and learning effects due to later positions in the test sequence were found by varying the order of subtests of an attention test battery (Földényi, Tagwerker-Neuenschwander, Giovanoli, Schallberger, & Steinhausen, 1999), thereby showing inconsistent results as a position later in the test sequence led to fatigue effects in some subtests but learning effects in other subtests. Tests that measure attention or memory seem to be particular vulnerable to fatigue effects, as they were also found in two subtests that measured memory and two subtests that involved attention by varying the sequence of a memory scale and an intelligence test battery (Zhu & Tulsky, 2000). But these effects were only found in analysis of the single subtest; they caused no main effects and the effect sizes were rather small. Learning effects were also found by varying two test parts of a test measuring field dependence-independence with figural item material (Kelleher, McRae, & Young, 1990); but the effects occurred only, when the test began with the more difficult part. Test order effects were found due to a randomized test sequence of tests that measured conceptual and procedural knowledge about decimal fractions which were interpreted as a lack of validity of these tests (Schneider & Stern, 2010).

Aim of the current experiment

Because of different reasons, such as organisational reasons (for example not all test are available on all computers), to avoid cheating, or to avoid a decrease of motivation if some test-takers realise that others are performing faster, the use of different test orders within test batteries is a common practice. But this practice is not well explored or proven to be without consequences for the test results, and there are barely studies that have investigated effects of test order regarding objective personality tests. Therefore, this paper presents an experiment with varied test orders with respect to so-called objective personality tests. Most previous studies were conducted with volunteers. Since it is of interest and of practical relevance to see how a subject's performance in actual selection situations is affected by test order, the sample of the current experiment consists of subjects within professional selection situations. It can be assumed that their motivation is consistent, as their goal was to perform well and thus to give their best.

Materials and Methods

Sample

The sample consists of 66 managers of an Austrian industrial corporation (an automotive supplier) in "higher management positions" (business managers, department chiefs, and team leaders), who attended an investigation of their professional potential (job-related cognitive ability and personality dimensions) to find out if they are suited for their position, if they should be given a position without managerial responsibility, or if they have the potential to obtain a higher position within the corporation. In this respect, the testing situation resembles a personnel selection situation as degradation to a position without managerial responsibility was within the bounds of possibility. Only four of the participants were female; the age of the participants varied from 26 to 54 years.

Measures

To identify the requirements the managers had to meet, a requirement analysis was conducted. According to the resulting requirements profile which comprised particular cognitive abilities, aspects of personality, and management styles, a test battery using the following tests was arranged.

Cognitive ability tests

Adaptive Matrices Test (AMT) – German edition (Hornke, Etzel, & Rettig, 2007). The AMT is a reasoning test consisting of items that fit the Rasch model and which are presented in an adaptive mode.

Intelligence Structure Test (IST 2000 R) – German edition (Liepmann, Beauducel, Brocke & Amthauer, 2007). The IST 2000 R is an intelligence test battery consisting of 11 subtests that measure verbal (3 subtests), numerical (3 subtests), and figural intelligence (3 subtests), as well as memory (2 subtests). For the current experiment, only the subtests measuring verbal intelligence and numerical intelligence were selected.

Objective Personality Tests

Objective personality tests sensu R. B. Cattell (e.g. 1958) are experiment-based assessments of behaviour which assess a personality construct by observing the subject's behaviour when working on a performance or ability task, while the observation and registration of the behaviour is done via computer (Kubinger, 2009b). They could be described as individual computerised assessments, which oftentimes include computer simulations of job-related tasks. According to selection settings, objective personality tests have the advantage of being less fakable than personality questionnaires (Baldinger, 2006; Kubin-

ger, 2009b), which, in contrast, are vulnerable to faking tendencies that particularly occur in real selection processes.

Work Styles – German edition (Kubinger & Ebenhöh, 2007). The test-battery consists of three subtests. Subtest 1 ("comparing area sizes") measures decisiveness, exactitude, and reflexivity, Subtest 2 ("coding symbols") measures proficiency level, aspiration level, target discrepancy, and frustration tolerance, and Subtest 3 ("distinguishing figures") measures achievement motivation.

Resilience-Assessment: computer based Objective Personality Test Battery (BAcO-D) – German edition (Ortner, Kubinger, Schrott, Radinger, & Litzenberger, 2006). BAcO-D is a test battery that measures different kinds of job-related resilience, and consists of six subtests. For the current experiment two of the six subtests were selected: the subtest "task collision" which measures one's resilience given multiple simultaneous tasks, and the subtest "crisscrossed plans" which measures one's resilience given thwarted plans.

ILICA – a simulation test to assess decisive behaviour – German edition (Möseneder & Ebenhöh, 1996). ILICA is a computer simulation test in German which measures self-management abilities. A leisure day is simulated for 30 minutes where arrangements for a nearing holiday have to be made.

Hypotheses

As certain context effects in questionnaires occur relating to the contents of prior questions, prior responses, or both (Smith, 1992), we hypothesize that similar context effects occur in test batteries due to test sequences. According to the findings of Fazio, Powell and Herr (1983) as well as La Rue and Olejnik (1980), we assume that the direct experience with different kinds of tests within a test battery might influence subjects' per-

formance in subsequent tests. If a subject works, for example, on computer simulations (objective personality tests) that provoke different kinds of stress or frustration to measure resilience, this might have an influence on his/her performance in subsequent cognitive ability tests, or vice versa. Thus, we expect that test performance varies depending on different test orders. We therefore formulate the following two hypotheses:

Hypothesis 1: The prior work on objective personality tests influences the subsequent performance in cognitive ability tests.

Hypothesis 2: The prior work on cognitive ability tests influences the subsequent performance in objective personality tests.

Design

Participants were randomly assigned to two experimental groups where the sequence of cognitive ability tests and objective personality tests was varied within a computer based test battery (see Figure 1). Experimental Group 1 completed objective personality tests first and cognitive ability tests subsequently (Test Order O), Experimental Group 2 completed cognitive ability tests prior to the objective personality tests (Test Order C). The cognitive ability tests were presented in a fixed order (AMT – IST 2000 R) as well as the objective personality tests (*Work Styles* – BAcO – ILICA).

Insert Figure 1 about here

Results

To investigate Hypotheses 1 and 2, a multivariate analysis of variance (MANOVA) was conducted to compare means in test scores between the two experimental groups in regard to the main factor test order. With $\alpha = .05$ and $\beta = .20$, a MANOVA is able to detect a mean difference of $\delta \geq 3/4\ \sigma$ (the standard deviation of the test scores) by testing 29x2 = 58 subjects. With 33x2 = 66 subjects, our sample size was designed adequately.

Results of the MANOVA

The means and standard deviations of all subtests in each experimental condition are given in Table 1. The two scores "target discrepancy" (*Work Styles*), and "target orientation" (ILICA) were deleted from the MANOVA successively as they caused inhomogeneities in the variance-covariance matrix and hence significant Box's M-Tests ($p = .044$; $p = .035$). Afterwards, the Box's M-Test proved to be not significant ($p = .063$). That is, the resulting *F*-values of the MANOVA can fairly be interpreted. The MANOVA showed a significant effect of test order ($p = .045$; $F = 1.842$; $\eta^2 = .450$). Considering the independently analysed test scores only two of them, "decisiveness" and "frustration tolerance", reveal significantly different means between the two experimental groups ($p = .00$ and $p = .041$). See the respective means in Table 1. In order to additionally investigate the factor test order on the scores "target discrepancy" and "target orientation", which were deleted from the MANOVA, two-sample *t*-tests for unequal variances (Welch tests) were applied: no significant effect occurred ($p = .480$, $p = .715$).

Insert Table 1 about here

Interpretation

According to MANOVA the results provide evidence so that Hypothesis 2 is to be accepted, but this is not true for Hypothesis 1. While test order had no effect on subjects' test scores in cognitive ability tests, test order influenced subjects' test scores in the objective personality test *Work Styles*. Subjects who worked on the *Work Styles* subsequently to two cognitive ability tests (AMT and IST 2000 R, which last approximately 2.5 hours after the very beginning of test administration) showed a significantly higher decisiveness insofar as they responded more quickly to the stimuli and they developed a significantly lower frustration tolerance as their willingness to perform decreased after a false negative feedback (cf. Table 1).

Discussion

The significant main effect of the factor test order in the MANOVA provides evidence in favour of Hypothesis 2 and against Hypothesis 1. Subjects who worked on ability tests first and on objective personality tests second showed higher *decisiveness* and lower *frustration tolerance* in the objective personality tests than subjects who worked on objective personality tests first. These effects resemble those of Baldinger (2006), who showed a similar effect of test order on the score "decisiveness" when the *Work Styles* were administered after personality questionnaires. Therefore, it seems that the test score "decisiveness" might be sensitive to test order effects irrespective of whether the *Work Styles* are administered after cognitive ability tests or personality questionnaires. Thus, this effect cannot be ascribed to the previously presented cognitive ability tests in particular. The increased decisiveness might possibly show a fatigue effect in the sense that subjects might have been tired due to the previous tests and therefore might have wanted to finish the testing more quickly, especially in regard to subtests that presented very simple tasks like

comparing area sizes or coding symbols in the *Work Styles*. However, test order did not affect challenging tasks, which are to be found in the other tests. It seems as if subjects tended to work faster depending on the test length, a phenomenon obvious from the score "decisiveness". A similar effect is described by Eiselt (1991), who found significant effects of test order on working time and test scores as working speed increased with progressing time while the number of correct answers decreased. In the current study, no differences in subjects' performance in the cognitive ability tests were revealed. Even if decisiveness increased, test order seems to have had no influence on the ability test scores (AMT, IST 2000 R).

Comparing our findings of the influence of test order on frustration tolerance to those of Hambros (2002), it seems very interesting that particularly contents potentially related to resilience show themselves to be vulnerable to order effects. The comparison of studies dealing with item order in personality questionnaires shows a similar picture, as the majority of order effects seem to be found foremost in questionnaires not mainly focussed on job- or achievement-related contents (cf. Knowles, 1988; Steinberg, 1994). Hence, one approach to explain our findings could be that test scores of personality questionnaires as well as test scores of objective personality tests assessing resilience seem to be more easily affected by test order than other test scores.

A second explanation for the lower frustration tolerance after working on cognitive ability tests might be the kind of frustration tolerance measured in the *Work Styles*. Frustration is provoked through a social comparison in the form of faked feedback (within a simple symbol coding task) that others have achieved higher scores. After receiving this feedback, subjects are asked to predict their achievements in the next level. Those subjects who reduce their achievement prediction after receiving the faked feedback are classified as showing a lower frustration tolerance. As all subjects are characterised by higher job

positions ("higher management"), as well as higher levels of education, we suggest that success and high achievements might have been important for them, and that frustration provoked through social comparisons in combination with the previously administered cognitive ability tests (Experimental Group 2) might have enhanced the frustration effect of the *Work Styles*. This hypothesis must, of course, be proved by further research.

In summary, it seems that cognitive ability test scores are less vulnerable to test order than objective personality test scores, or that the observed effects of test order occur only in tests or subtests with very simple tasks while more complex tasks are not affected. This effect is somewhat similar to the findings of Hambros (2002), who reported that previously presented cognitive ability tests had an influence on subsequent personality questionnaires while there was no reverse effect.

Limitations and implications for future research

The current study has a limitation that future research should address. The sample is very special as most participants were men (only four of the participants were female), and the participants were all managers in "higher management positions" of a particular industrial corporation (an automotive supplier). Therefore, the findings might not apply to other groups, such as women, other occupational groups, or other positions within similar professions. Although different test orders affected some of the tests scores, most test scores were not affected. Nevertheless, the revealed effects seem to be more complex and should be further explored. A comprehension of the possible interactions between sample, task type, content of the measurement, and the modality of specific simulations might reveal explanations for context effects in other measurements. It might also be interesting to find out more about what kind of simulations or tasks are vulnerable to context effects, especially in regard to the measurement of resilience. The current paper revealed that some

kinds of tasks measuring resilience were affected by varied test order (*Work Styles*) while others were not (BAcO-D). The reasons might be the above mentioned type of frustration provoked through social comparison, or the difficulty or complexity of tasks. Further research should take this into consideration more precisely.

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology, 15*, 163-181.

Baehr, M. E. (1953). A simplified procedure for the measurement of employee attitudes. *The Journal of Applied Psychology, 37*, 163-167.

Baldinger, D. (2006). *Der Einfluss sozial erwünschten Verhaltens auf das Ergebnis Objektiver Persönlichkeitstests [The influence of social desirable behaviour on the results of objective personality questionnaires]*. Unpublished diploma thesis, University of Vienna.

Cattell, R. B. (1958). What is "objective" in "objective" personality tests? *Journal of Counseling Psychology, 5*, 285-289.

Eiselt, W. (1991). *Der Einfluss subjektiver Daten und Testreihenfolge auf die Leistung in einer computergestützten Testbatterie [The influence of subjective data and test order on the achievement within a computer based test battery]*. Investigation of the psychological office of the German armed forces (Navy Medical Institute of the Marine), Kiel, 5-54.

Eisenhauer, E. (2008). *Effekte bei der Veränderung der Itempositionen anhand des Anstrengungsvermeidungstest und B5PO [Effects of different item positions by means*

of the exertion avoidance test and B5PO]. Unpublished thesis, University of Vienna.

Fazio, R. H., Powell, M. C., & Herr, P. M. (1983). Toward a process model of the attitude-behavior relation: Accessing one's attitude upon mere observation of the attitude object. *Journal of Personality and Social Psychology, 44,* 723-735.

Földényi, M., Tagwerker-Neuenschwander, F., Giovanoli, A., Schallberger, U., & Steinhausen, H.-C. (1999). Die Aufmerksamkeitsleistungen von 6-10-jährigen Kindern in der TAP [Attentional performance of 6-10-year-old children on the TAP]. *Zeitschrift für Neuropsychologie, 10,* 87-102.

Franke, G. H. (1997). „The whole is more than the sum of its parts": The effects of grouping and randomizing items on the reliability and validity of questionnaires. *European Journal of Psychological Assessment, 13*, 67-74.

Hambros, K. (2002). On reasonableness of personality inventories with dichotomous item response format. *Psychologische Beiträge* [latterly: *Psychological Test and Assessment Modeling*]*, 44,* 126-135.

Hartig, J., Hölzel, B., & Moosbrugger, H. (2007). A confirmatory analysis of item reliability trends (CAIRT): Differentiating true score and error variance in the analysis of item context effects. *Multivariate Behavioral Research, 42,* 157-183.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*, 373-385.

Hornke, L. F., Etzel, S., & Retting, K. (2007). *Adaptiver Matrizentest (AMT) [Adaptive Matrices Test]*. Version 26.00, Mödling: Schuhfried.

Kelleher, W. E., McRae, L. S. E., & Young, J. D. (1990). The group embedded figure test: The learning effect reexamined. *Perceptual and Motor Skills, 70,* 1233-1234.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*, 312-320.

Knowles, E. S., & Byers, B. (1996). Reliability shifts in measurement reactivity: Driven by content engagement or self-engagement? *Journal of Personality and Social Psychology, 70*, 1080-1090.

Knowles, E. S., Coker, M. C., Cook, D. A., Diercks, S. R., Irwin, M. E., Lundeen, E. J., Neville, J. W., & Sibicky, M. E. (1992). Order effects within personality measures. In N. Schwarz & S. Sudman. *Context effects in Social and Psychological Research,* (pp. 221-236). New York: Springer.

Kraut, A. I., Wolfson, A. D., & Rothenberg, A. (1975). Some effects of position on opinion survey items. *Journal of Applied Psychology, 60*, 774-776.

Kubinger, K. D. (2009a). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement, 69*, 232-244.

Kubinger, K. D. (2009b). The technique of objective personality-tests sensu R. B. Cattell nowadays: The Viennese pool of computerized tests aimed at experiment-based assessment of behaviour. *Acta Psychologica Sinica*, *41*, 1024-1036.

Kubinger, K. D., & Ebenhöh, J. (2007). *Arbeitshaltungen (AHA) [Work Styles]*. Version 27.00, Mödling: Schuhfried.

La Rue, A., & Olejnik, A. B. (1980). Cognitive "priming" of principled moral thought. *Personality and Social Psychology Bulletin, 6,* 413-416.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55,* 387-413.

Liepmann, D., Beauducel, A., Brocke. B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R (IST 2000 R) [Intelligence Structure Test]*. 2. revised edition, Göttingen: Hogrefe.

Möseneder, D., & Ebenhöh, J. (1996). *Ein Simulationstest zur Erfassung des Entscheidungsverhaltens (ILICA) [A simulation test for the measure of decision behaviour]*. Frankfurt: Swets.

Ortner, T. M. (2008). Effects of changed item order: A cautionary note to practitioners on jumping to computerized adaptive testing for personality assessment. *International Journal of Selection and Assessment, 16*, 249-257.

Ortner, T. M., Kubinger, K. D., Schrott, A., Radinger, R., & Litzenberger, M. (2006). *Belastbarkeits-Assessment: computerisierte Objektive Persönlichkeits-Testbatterie – Deutsch (BAcO-D) [Resilience-Assessment: computer based Objective Personality Test Battery – German version]*. Frankfurt/M.: Harcourt Test Services.

Perlini, A. H., Lind, D. L., & Zumbo, B. D. (1998). Context effects on examinations: The effects of time, item order and item difficulty. *Canadian Psychology/Psychologie canadienne, 39*, 299-307.

Rost, D. H., & Hoberg, K. (1996). *Itempositionsveränderungen in Persönlichkeitsfragebogen: Methodischer Kunstfehler oder tolerierbare Praxis? [Changing the position of items in personality questionnaires: Methodological malpractice or tolerable practice?]*. Reports of the Department of Psychology, Vol. 114, Philipps-University Marburg.

Schneider, M., & Stern, E. (2010). The developmental relations between conceptual and procedural knowledge: A multimethod approach. *Developmental Psychology, 46,* 178-192.

Schriesheim, C. A. (1981). Leniency effects on convergent and discriminant validity for grouped questionnaire items: A further investigation. *Educational and Psychological Measurement, 41,* 1093-1099.

Schriesheim, C. A., Kopelman, R. E., & Solomon, E. (1989). The effect of grouped versus randomized questionnaire format on scale reliability and validity: A three-study investigation. *Educational and Psychological Measurement, 49,* 487-508.

Schwarz, N., & Wyer, R. (1983). Effects of rank ordering stimuli on magnitude ratings of these and other stimuli. *Journal of Experimental Social Psychology, 21,* 30-46.

Smith, T. W. (1992). Thoughts on the nature of context effects. In N. Schwarz & S. Sudman. *Context effects in Social and Psychological Research,* (pp. 163-184). New York: Springer.

Sparfeldt, J. R., Schilling, S. R., Rost, D. H., & Thiel, A. (2006). Blocked versus randomized format of questionnaires – A confirmatory multigroup analysis. *Educational and Psychological Measurement, 66,* 961-974.

Steinberg, L. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology, 66,* 341-349.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103,* 299-314.

Zhu, J., & Tulsky, D. S. (2000). Co-norming the WAIS-III and WMS-III: Is there a test-order effect on IQ and memory scores? *The Clinical Neuropsychologist, 14*, 461-467.

Table 1:
Means and standard deviations for all subtests in each experimental condition (2 experimental groups) with regard to the factor "test order"

Dependent variable (test score)	Test order O		Test order C	
	Mean	Standard deviation	Mean	Standard deviation
reasoning – AMT	.445	.787	.447	.899
verbal intelligence – IST 2000 R	.155	.898	.277	.975
numerical intelligence – IST 2000 R	.536	1.000	.549	.957
reflexivity – *Work Styles*	.486	.923	.114	1.130
exactitude – *Work Styles*	.428	1.148	.594	.946
decisiveness – *Work Styles*	-.558	.651	.209	.776
proficiency level – *Work Styles*	.110	.673	.251	.534
aspiration level – *Work Styles*	-.014	.723	.052	.857
frustration tolerance – *Work Styles*	.595	.854	.205	.651
target discrepancy – *Work Styles*	-.4269	.74599	-.3105	1.27918
main task – BAcO-D task collision	-.563	1.370	-.547	1.274
efficiency of subsidiary task – BAcO-D task collision	-.184	.936	-.227	.800
quantity of subsidiary task – BAcO-D task collision	-.013	1.174	-.003	1.058
perseverance – BAcO-D task collision	-.521	1.109	-.824	1.159
balance – BAcO-D task collision	-.019	1.703	-.056	1.626
quantity – BAcO-D crisscrossed plans	2.359	2.264	2.624	2.145
speed – BAcO-D crisscrossed plans	-.596	.772	-.472	.676
abidance – BAcO-D crisscrossed plans	.931	.420	.936	.507
stray – BAcO-D crisscrossed plans	.460	.995	.114	.847
flexibility – ILICA	-.289	.855	.452	.793
target orientation – ILICA	.871	1.013	.834	.789
distractibility – ILICA	-.022	1.000	.068	.778

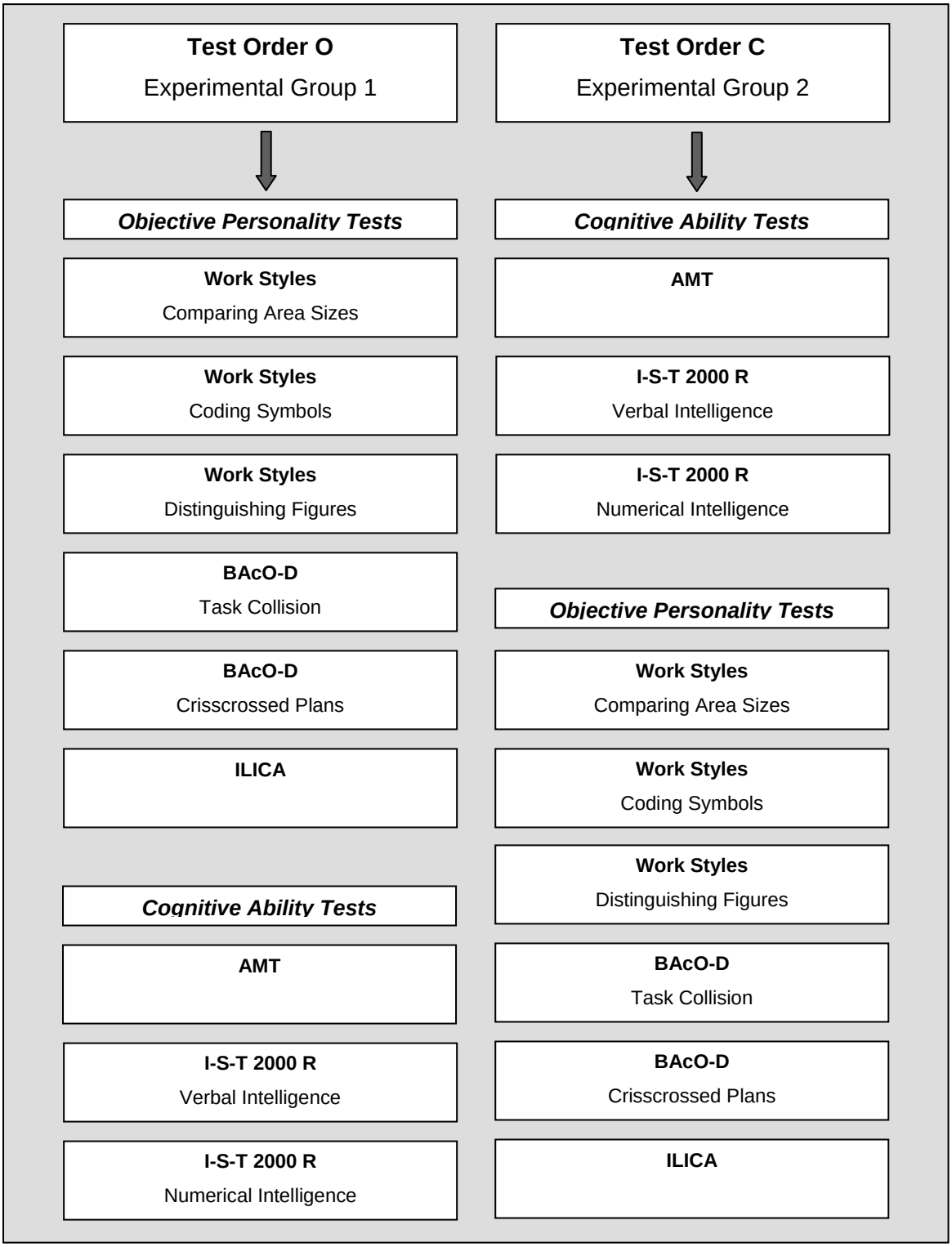


Figure 1: Experimental design

6. Tables and Additional Papers

(Appendix 1 – 2)

6.1. Appendix 1

1) Table 4, Paper 1 (Khorramdel & Kubinger, 2006): complete version
Means and standard deviations (SD) for all scales by experimental condition
(8 experimental groups)

2) Table 1, Additional Study
Means and standard deviations for all scales in each experimental condition
(8 experimental groups) with regard to the factors Response Format, Response Time, and Instruction

Table 4 (Paper 1):

Means and standard deviations (SD) for all scales by experimental condition (8 experimental groups)

	No Warning				Warning			
	Dichotomous		Analogue		Dichotomous		Analogue	
Scale	No Lim.	Lim.	No Lim.	Lim.	No Lim.	Lim.	No Lim.	Lim.
extroversion	16.77	14.86	13.15	16.30	15.41	15.43	15.66	14.68
	(5.84)	(6.03)	(4.40)	(5.24)	(4.55)	(7.26)	(5.15)	(6.58)
introversion	7.82	9.67	11.55	8.55	9.29	9.43	8.88	9.68
	(5.71)	(5.65)	(4.15)	(4.89)	(4.36)	(6.79)	(4.97)	(6.57)
sensing	10.05	10.05	12.50	11.40	11.65	9.95	10.88	10.68
	(3.51)	(4.82)	(4.03)	(3.17)	(4.25)	(4.91)	(4.08)	(5.94)
intuition	7.14	7.57	6.50	6.35	6.33	7.67	7.25	7.58
	(2.98)	(3.63)	(2.96)	(2.56)	(3.06)	(3.92)	(3.37)	(4.43)
thinking	10.45	9.57	10.00	11.05	9.43	9.43	8.53	9.47
	(2.92)	(4.37)	(2.25)	(2.50)	(3.53)	(2.68)	(3.11)	(3.01)
feeling	6.95	7.52	7.25	6.10	7.61	7.86	8.47	8.11
	(2.57)	(3.50)	(1.92)	(2.36)	(3.02)	(2.13)	(2.86)	(2.47)
judging	10.73	11.19	10.55	10.70	10.63	10.19	10.00	10.89
	(2.99)	(3.46)	(2.82)	(2.56)	(3.30)	(3.68)	(3.21)	(3.43)
perceiving	6.14	6.24	7.10	6.15	6.75	7.10	7.22	6.32
	(3.96)	(3.83)	(3.67)	(3.57)	(4.02)	(4.33)	(4.29)	(4.22)
self-concept of own comp.	1.18	2.76	2.65	1.40	2.25	3.05	2.75	1.68
	(1.87)	(2.43)	(1.81)	(1.27)	(2.07)	(3.13)	(1.97)	(1.70)
internality	2.14	1.90	2.15	1.65	2.43	2.71	1.72	1.89
	(1.36)	(1.22)	(1.35)	(1.27)	(1.72)	(1.42)	(1.28)	(1.56)
powerful others control	6.27	5.10	4.60	5.70	5.53	5.57	4.72	5.53
	(1.49)	(2.05)	(2.06)	(1.63)	(2.10)	(1.83)	(1.90)	(2.20)
chance control	6.27	5.00	3.55	5.80	5.57	3.67	4.47	5.79
	(1.83)	(2.21)	(2.04)	(1.74)	(1.79)	(2.50)	(2.14)	(1.58)

Note. Within each row, means are presented above and standard deviations presented below in parentheses.

Table 1 (Additional Study):

Means and standard deviations for all scales in each experimental condition

(8 experimental groups)

	No Warning				Warning			
	Dichotomous		Analogue		Dichotomous		Analogue	
Scale	No Lim.	Lim.	No Lim.	Lim.	No Lim.	Lim.	No Lim.	Lim.
extroversion	19.90	16.60	16.10	16.80	15.57	18.63	17.23	18.60
	(4.17)	(5.64)	(2.84)	(5.25)	(4.80)	(4.47)	(4.01)	(3.02)
introversion	4.50	8.50	7.80	8.00	9.00	6.18	7.33	5.60
	(3.74)	(5.64)	(2.89)	(4.98)	(4.47)	(4.16)	(3.66)	(2.87)
sensing	11.20	9.90	8.90	11.46	12.74	9.18	12.80	9.90
	(3.01)	(4.48)	(3.90)	(3.46)	(4.35)	(5.15)	(3.84)	(5.70)
intuition	6.30	7.60	9.00	6.66	5.71	9.18	5.95	8.60
	(2.35)	(3.13)	(2.49)	(2.69)	(2.87)	(3.60)	(3.07)	(4.16)
thinking	11.20	11.40	11.00	11.06	10.11	8.90	9.19	8.90
	(2.85)	(2.83)	(5.18)	(2.46)	(3.37)	(3.70)	(3.18)	(2.37)
feeling	7.70	6.70	8.10	7.53	8.20	9.90	9.23	10.00
	(2.83)	(2.45)	(4.55)	(2.23)	(3.03)	(3.30)	(2.84)	(2.05)
judging	11.50	9.80	11.10	11.20	11.05	9.00	11.19	10.80
	(2.12)	(3.55)	(3.07)	(2.27)	(3.10)	(2.64)	(3.02)	(3.01)
perceiving	5.50	7.60	6.80	5.20	6.31	8.81	5.66	6.30
	(3.02)	(3.59)	(3.91)	(2.90)	(3.78)	(3.09)	(3.73)	(3.77)
self-concept of own comp.	.50	1.50	2.00	1.13	2.37	1.72	2.57	1.20
	(.70)	(1.77)	(1.94)	(1.12)	(2.19)	(2.00)	(2.03)	(1.03)
internality	1.70	2.10	1.10	1.53	2.74	2.00	1.66	1.90
	(.48)	(1.59)	(1.10)	(1.30)	(1.66)	(1.73)	(1.35)	(1.44)
powerful others control	6.40	5.10	5.80	5.80	5.77	6.00	5.00	5.50
	(1.17)	(2.42)	(1.98)	(1.37)	(2.17)	(1.73)	(1.84)	(1.84)
chance control	7.10	5.60	5.70	5.93	5.71	5.81	5.04	6.30
	(.87)	(1.83)	(2.05)	(1.43)	(1.80)	(1.53)	(1.96)	(1.82)

Note. Within each row, means are presented above and standard deviations presented below in parentheses.

6.2. Appendix 2

1) Paper 4:

Khorramdel, L., & Kubinger, K. D. (submitted). The influence of different rating scales on impression management. Should we give um on rating scales? *Journal of Personality Assessment.*

2) Paper 5:

Khorramdel, L., & Frebort, M. (submitted). Test order effects: a one-hit-wonder? Trying to replicate the findings of Khorramdel and Frebort (2010). *European Journal of Psychological Assessment*.

The influence of different rating scales on impression management

Should we give up on rating scales?

Lale Khorramdel and Klaus D. Kubinger
University of Vienna

Lale Khorramdel, Klaus D. Kubinger, and Alexander Uitz Center of Testing and Consulting, Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna.
Correspondence concerning this article may be addressed to Lale Khorramdel, University of Vienna, Liebiggasse 5, 1010 Vienna, Austria, Email : lale.khorramdel@univie.ac.at, or Klaus Kubinger, Email: klaus.kubinger@univie.ac.at

Abstract

The influence of different rating scales on socially desirable response distortion was investigated by administering the Personality Research Form (PRF; Stumpf, Angleitner, Wieck, Jackson, & Beloch-Till, 1984) to 268 applicants with either a 6-point rating scale or a 2-point rating scale. It was hypothesised that a 6-point rating scale leads to less response distortion than a 2-point rating scale as it might be more difficult to adjust responses to a faking good schema. Results provide evidence of the advantages of the 6-point rating scale and show that the type of response format might interact with item content and wording.

Key words: faking good, impression management, personnel selection, response format, social desirability

Introduction

The investigation of intentional response distortions by giving socially desirable or job-related desirable answers (faking good or impression management) has kept a lot of researchers busy over the past few years (cf. Rothstein & Goffin, 2006). Meanwhile an immense number of studies dealing with this topic have appeared. One reason for the growing interest in this topic might be the increasing interest from organizations to include personality questionnaires in their selection processes in order to find the most suitable applicants. Even if personality measures seem to have a low validity for predicting overall job performance, they may improve the validity of the selection process when combined with cognitive ability tests (Morgeson, Campion, Diboye, Hollenbeck, Murphy, & Schmitt, 2007a, 2007b) – of course, there is also the opinion that published personality tests yield useful validity estimates (Ones, Dilchert, Viswesvaran, & Judge, 2007; Tett & Christiansen, 2007). Personality questionnaires are favoured measures in this respect because they complement other selection methods which measure the "capacity to perform aspect" (or can do aspect) by measuring the "choice to perform aspect" (or will do aspect) of job performance (Goffin & Boyd, 2009). But personality questionnaires seem to be particularly vulnerable to response distortions (Ziegler, Schmidt-Atzert, Bühner, & Krumm, 2007) because of their high transparency that often makes the measured construct evident to the test-taker (Furnham, 1986), who may in turn have a proclivity to present himself in a socially desirable way or in a way that is desirable with regard to the aspired job. It was in actual fact shown that personality questionnaire scales were affected by applicants' response distortion while objective personality tests (experiment-based assessments of behaviour) were not affected, maybe because the computation of the test scores is not transparent to test-takers (Ziegler, Schmidt-Atzert, Bühner, & Krumm, 2007).

Some individuals are more capable of faking than others which leads to the assumption that fakability is a variable of individual differences (Mueller-Hanson,

Heggestad, & Thornton, 2006; Viswesvaran, & Ones, 1999), and individuals who have knowledge of the aspired job and the appropriate desired behaviours might be able to fake their responses better (Goffin & Boyd, 2009; Levashina & Campion, 2006; Snell, Sydell, and Lueke, 1999). According to different variables that moderate faking behaviour (see also different models of impression management below) this behaviour occurs with different faking styles, for instance slight versus extreme faking (Robie, Brown, & Beaty, 2007; Zickar, Gibby, & Robie, 2004). Different models of impression management were proposed to explain the process underlying faking so as to support further research, and strategies were investigated in order to find possibilities of dealing with faking or suppressing it. Both efforts are described in the following. But first, the usefulness of personality questionnaires in personnel selection and the need for further research is discussed.

Usefulness of personality questionnaires in personnel selection

A lively discussion about the usefulness of personality questionnaires in selection processes has been ongoing between researchers. Some researchers assume that socially desirable response distortions could be an intelligent adaptation to situations, socially adaptive, or an expression of social competence that could predict job performance (Marcus, 2003a, 2003b; Morgeson, Campion, Diboye, Hollenbeck, Murphy, & Schmitt, 2007a), a hypothesis not yet proven (Jackson Foldes, Ones, & Sinangil, 2006; Ones, Viswesvaran, & Reiss, 1996; Viswesvaran, Ones, & Hough, 2001). It was shown rather that measures of self-monitoring and social desirability (impression management and self-deception) did not function as performance predictors (Li & Bagger, 2006) or as predictor variables with respect to other criterion measures (e.g. interviews, role plays, group discussions, work samples, cognitive ability tests, in-tray exercise), but did also not serve as confounding variables (Li & Bagger, 2006). It has also been argued that intentional

response distortion cannot serve as a common predictor for job performance as applicants use different strategies and tactics of self-description, and that not every strategy or tactic is successful in every job (Kanning, 2003). Moreover, it has not yet been investigated, whether the use of fakable measures in personnel selection affects an organisation's image, which in turn could have consequences on the availability of applicants or the attitude of incumbents towards the organisation (Kersting, 2004). Maybe an interesting finding with respect to the idea of using socially desirable response distortions to predict job performance is that social desirability (measured with a social desirability scale) seems to be related to self-esteem and emotional intelligence as they shared common variance, and that over-claiming is not correlated with social desirability, at least in a volunteer sample that is not comparable to samples in high-stakes assessments (Mesmer-Magnus, Viswesvaran, Deshpande, & Joseph, 2006).

There are studies (mostly meta-analyses) which provide evidence that intentional response distortions barely affect construct validity or criterion-related validity of self-reported personality measures (Bradley & Hauenstein, 2006; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Ellingson, Smith, & Sackett, 2001; Moorman & Podsakoff, 1992; Ones, Dilchert, Viswesvaran, & Judge, 2007; Ones, Viswesvaran & Reiss, 1996; Ones, Viswesvaran & Schmidt, 1993; Smith & Ellingson, 2002). Other studies have been able to show that criterion-related validity is affected at the high end of the predictor distribution and that rank order changes take place influencing which applicant gets hired (Ellingson, Sackett, & Hough, 1999; Griffith, Chmielowski, & Yoshita, 2007; Mueller-Hanson, M., Heggestad, E. D., & Thornton, G. C., 2003; Robie, Brown, & Beaty, 2007; Rosse, Stecher, Miller, & Levin, 1998; Winkelspecht, Lewis, & Thomas, 2006). There are also findings that show slightly lower reliabilities of the questionnaire scales in applicant samples than in non-applicant samples, and it was also shown that the underlying constructs of a questionnaire were measured differently across these samples harming the construct

validity (Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001). Hence, some researchers assume social desirability to be a serious problem if self-reported personality scales are used in selection (Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001; Winkelspecht, Lewis, & Thomas, 2006).

Need for further research

If intentional response distortions in personality questionnaires do harm selection processes by preventing the selection of the appropriate applicants, alternative measures that are less susceptible should be used (Morgeson, Campion, Diboye, Hollenbeck, Murphy, & Schmitt, 2007a; Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001), or response distortions should be suppressed or controlled. Even if faking would not decrease the usefulness of personality testing in personnel selection, their value could be increased if more would be known about the process underlying faking (Goffin & Boyd, 2009). Models which try to define the underlying process are helpful but yet not fully developed and previously investigated strategies to deal with faking have revealed inconsistent results or have not permitted any general conclusions. Therefore, further studies that investigate different aspects of faking could provide helpful contributions.

In addition, the investigation of socially desirable response distortion could provide an important contribution to personality research and social theories.

Models of impression management

Models, which try to explain the process which underlies intentional response distortion, describe faking behaviour as an interaction of the ability and motivation to fake (Snell, Sydell, & Lueke, 1999), the opportunity to fake depending on the test-takers true score (McFarland & Ryan, 2000), the perception of the situation (belief in the importance of faking, perceived behavioural control and beliefs about subjective norms), and general personality characteristics like Conscientiousness and Emotional stability (McFarland &

Ryan, 2006; Mueller-Hanson, Heggestad, & Thornton, 2006). A further model expanded these aspects by suggesting that faking is likely to vary depending on the specific nature of the personality item (instead of considering the entire personality test), and including the factor "perceived ability to fake" which is more relevant to the motivation or intention to fake than the true ability to fake (Goffin & Boyd, 2009).

Regarding the information processing, it is suggested that intentional response distortions rely on the comparison of the item content with an adopted schema (instead of a self-schema) like a schema of favourable impressions (Holden & Hibbs, 1995; Holden, Kroner, Fekken, & Popham, 1992). Therewith, a systematic relationship between the presence of a schema and response time latencies is proposed. Thus, schema-congruent responses (e.g. socially desirable responses by adopting a fake good schema) are given faster than schema-incongruent responses (e.g. socially undesirable responses with regard to a fake good schema). But the resulting response latencies may also be influenced by item characteristics (item length, item ambiguity, item extremity, number of alternatives) and multiple person variables (reading speed, verbal ability, motor speed) (see also Holden, Fekken, & Cotton, 1991). Furthermore, it was shown that test-takers who were instructed to fake good on a personality questionnaire used job stereotypes (or schemas) however with negative aspects removed, as well as stereotypes about general aspects of personality (Mahar, Coburn, Griffin, Hemeter, Potappel, Turton, & Mulgrew, 2006; Mahar, Cologon, & Duck, 1995). The latter was assumed because training test-takers had no impact on the use of stereotypes in contrast to test-takers who were not trained. These findings conflict with the assumption that the more knowledge persons have about the aspired job the easier it is for them to fake (Goffin & Boyd, 2009; Levashina & Campion, 2006; Snell, Sydell, and Lueke, 1999).

As even a small degree of the ability to fake combined with the motivation to do so, as well as a small degree of motivation to fake combined with the ability to fake can

already lead to response distortion, an approach to reduce faking might be more successful if both ability and motivation to fake are considered (Rothstein & Goffin, 2006). Such an approach was attempted by Khorramdel and Kubinger (2006; see below).

Strategies to deal with faking good

Strategies to deal with faking good, like the identification of response distortion (with the use of response time latencies or social desirability scales), the discouragement of test-takers from faking (with warning instructions that faking can be identified or will have negative consequences), or efforts to make personality questionnaires less fakable (by adjusting the response format, method of administration, or item positioning) have provided inconsistent results, and no general conclusions can be drawn for these adjustments as the effects may be influenced by many moderating variables (Dilchert, Ones, Viswesvaran, & Deller, 2006). Strategies to make personality questionnaires less fakable have focused on aspects of how questionnaires are administered by adjusting the method of administration, the item positioning or the response format. In the following, strategies that imply the use of different response formats or that focus on item content shall be described in more detail as they are of interest for the current study.

Item content: Covert item contents or particular dimensions that have some positive as well as some negative aspects like extraversion (in contrast to dimensions that have almost only negative associations like neuroticism) are assumed to be more difficult to fake (Furnham, 1986). Reducing specific contents of items in personnel selection which are particularly relevant to a job (like reducing the college relevance of items within a college selection process) are supposed to make the items less fakable (Ramsay, Schmitt, Oswald, Kim, & Gillespie, 2006).

Response format: In contrast to items with a single-stimulus response format (rating scale), items with a forced-choice response format are assumed to minimize faking

tendencies, as they make it more difficult to respond desirably (Jackson, Wroblewski, & Ashton, 2000). These types of response formats may influence the perceived opportunity to fake, which in turn moderates the actual faking behaviour (Goffin & Boyd, 2009; Rothstein & Goffin, 2006; Snell, Sydell, & Lueke, 1999).

Instructing test-takers to respond like applicants (fake good conditions), less faking tendencies were revealed by using binary and quartet forced-choice formats, as well as ipsative and partially ipsative forced-choice formats, while single-stimulus response formats were more vulnerable to response distortions (Jackson, Wroblewski, & Ashton, 2000; Martin, Bown, & Hunt, 2002). Nevertheless, it has been shown that test-takers are able to distort their responses using a forced-choice format (Lammers & Frankenfeld, 1999). The forced-choice format was additionally shown to be a better predictor of personality and job-related abilities in fake good conditions than the single-stimulus format (Jackson, Wroblewski, & Ashton, 2000; Wright & Miederhoff, 1999). It was also revealed that both types of response formats were susceptible to response distortions, that subjects with higher cognitive ability were able to distort their answers more by using the forced-choice format than subjects with lower cognitive ability, and that a forced-choice response format was not better at retaining the rank ordering of individuals in comparison to a single-stimulus response format (Christiansen, Burns, & Montgomery, 2005; Heggestad, Morrison, Reeve, & McCloy, 2006). However, items with the forced-choice format showed higher construct validity under fake good conditions (Christiansen, Burns, & Montgomery, 2005). With respect to criterion-related validity no difference between these two response formats was revealed (Converse et al., 2008). As the samples in these studies consisted of volunteers (mostly students) the results still need to be demonstrated in real selection situations.

Some research suggests that using analogue scales (in which participants mark the extent of their agreement along a continuous line between two alternatives) as a response

format may be less prone to faking than a dichotomous (participants have to choose one of two alternatives) response format (Khorramdel & Kubinger, 2006; Kubinger, 2002; Seiwald, 2002). It has also been suggested that a dichotomous response format provokes a kind of reactance resulting in untypical or arbitrary responses that do not describe the subjects' true character (Karner, 2002).

A combination of different approaches shown by Khorramdel and Kubinger (2006) revealed not only the advantage of an analogue scale in contrast to a dichotomous response format (main effect of the factor response format in MANOVA), but also interaction effects of the factors response format and speed. A limited response time (per questionnaire page) led to less socially desirable responses than no time limit but only in combination with an analogue scale, while the reverse effect occurred by using a dichotomous forced-choice format. Considering the fact that single-stimulus (or normative) items as well as forced-choice items were used, either presented with an analogue scale or dichotomous (and regarding the findings about forced-choice response formats in other studies) it seems interesting that the effects occurred only in questionnaire scales with single-stimulus items, while no effects occurred in scales with forced-choice items. It might therefore be assumed that items with a single-stimulus response format (or rating scale) are less susceptible to faking when presented as an analogue scale rather than dichotomously. However, these two presentation types did not affect response behaviour when items were of a forced-choice-type format.

Aim of the current experiment

As single-stimulus items are mostly used in personnel selection because of their normative qualities, and because fewer costs are incurred with respect to their construction (compared to forced-choice formats), one aim of the current experiment is to find out if the susceptibility of items with a single-stimulus response format can be decreased by using

particular presentation modes. Another aim is to provide a contribution to personality research and social theories by investigating the interaction of variables that might moderate response behaviour. The current paper therefore investigates the effects of different presentation types of items with a single-stimulus response (or normative) format on socially desirable response distortions in personnel selection. The normative items are presented either with a 6-point rating scale (with 1 = disagree totally to 6 = agree completely) or dichotomously with a 2-point rating scale (with 1 = wrong and 2 = right).

Hypothesis

According to the findings of Khorramdel and Kubinger (2006) we expect that, similar to an analogue scale, the 6-point rating scale is less vulnerable to socially desirable response distortion than a 2-point rating scale (when normative items are used) as it is more difficult to figure out what degree of agreement is socially desired. Additionally, both the 6-point as well as the 2-point rating scales do not allow evasive responses as they have no middle category. Furthermore, we assume that test-takers' ability to adjust their responses in a questionnaire to an adopted faking good schema (Holden & Hibbs, 1995; Holden, Kroner, Fekken, & Popham, 1992) increases when a 2-point rating scale is used because it is easier to give stereotypical responses, while a 6-point rating scale might force test takers to consider their responses more precisely. Therefore, we formulate the following hypothesis:

Hypothesis 1: Normative items presented with a 6-point rating scale are less susceptible to socially desirable response distortion than with a 2-point rating scale.

Additionally, we formulate the null hypothesis and a second alternative hypothesis:

Hypothesis 2: Normative items presented with a 6-point rating scale are not less susceptible to socially desirable response distortion than with a 2-point rating scale.

Hypothesis 3: Normative items presented with a 2-point rating scale are less susceptible to socially desirable response distortion than with a 6-point rating scale.

Method

An experiment within a selection procedure is presented to investigate the effects of different response formats of normative items (dichotomous rating scale versus rating scale with 6 response categories) on faking personality questionnaires. The sample consists of soldiers in the Austrian Federal Armed Forces who had applied for a pilot training and underwent a selection procedure.

Sample

All applicants were soldiers of the Austrian Federal Armed Forces who had applied for pilot training. The sample consists of 84 applicants from another experiment, to which we refer as Study 1 in the following, and 184 applicants who were tested in the customary selection setting of the Federal Armed Forces. The current experiment is therefore a quasi experiment as the applicants were not randomly assigned to the experimental groups, however it can be assumed that there are no considerable differences between the applicants as they had applied all for the same training and come from the same institution. Furthermore, all applicants had undergone a pre-selection process in the Austrian Federal Armed Forces, before they applied for the training. In the following we refer to this quasi experiment as Study 2. The applicants of Study 1 filled out a paper-pencil questionnaire with a rating scale with six categories. The applicants of Study 2 filled out 234 items of the same questionnaire but this time a computer based version with a dichotomous rating scale. The data of the 184 applicants (Study 2) were provided by the Department of Human Resources in the Austrian Federal Armed Forces. Their first language is German, their education levels varies from a compulsory education of 9 years and apprenticeship to

general qualification for university entrance. Finally, the sample consists of a total of 268 applicants.

Measures

The following personality questionnaire was already a fixed part of the selection procedure in the Department of Human Resources and therefore used in Study 2, in order to investigate the effect of response-type formats on intentional response distortion:

Personality Research Form (PRF) – German edition. The PRF (Stumpf, Angleitner, Wieck, Jackson, & Beloch-Till, 1984) is a personality questionnaire based on Murray's personality theory (1938) that measures a set of traits important for psychological research as well as psychological assessment. 234 items with a dichotomous response format ("right" and "wrong") measure fifteen scales: Achievement, Affiliation, Aggression, Dominance, Endurance, Exhibition, Harm avoidance, Impulsivity, Nurturance, Order, Play, Social Recognition, Succorance, Understanding. The cronbachs alpha reliabilities of the PRF scales, presented in the test-manual, range from .69 to .87.

Design

Experimental group 1: The 84 applicants from Study 1 received the 234 PRF items with a 6-point rating scale (with 1 = disagree totally to 6 = agree completely) embedded in a paper-pencil questionnaire that also comprised 210 items from the BIP (Business-focused Inventory of Personality, German edition; Hossiep & Paschen, 2003), 60 items from the NEO FFI (NEO Five Factor Inventory, German edition; Borkenau & Ostendorf, 1993), and 12 items from the NEO PI-R (NEO Personality Inventory Revised, German edition; Ostendorf & Angleitner, 2004). This questionnaire was administered with two different item orders (with respect to the items from the BIP, NEO FFI, and NEO PI-R), whereas the positions of the PRF items were always held constant (in the middle of the questionnaire).

Experimental group 2: The 184 additional applicants for the current Study 2 received a computer version of the PRF with a dichotomous or 2-point rating scale (with 1 = wrong and 2 = right), and without the items from the other questionnaires.

Procedure

First, all applicants attended a psychological assessment carried out by the department of Human Resources of the Austrian Federal Armed Forces, where their cognitive abilities and personality traits were tested. After an assessment of approximately 8 to 9 hours (including cognitive ability and achievement tests), they filled out either the paper-pencil questionnaire by responding to a rating scale with six categories, or the computer based questionnaire with a dichotomous rating scale. None of the applicants had received any information about the requirements profile. It was not possible to randomly assign the applicants to the two experimental groups, but it can be assumed that there are no nameable differences between the applicants (see the description in the text "Sample" above). To be able to compare the two experimental groups in our analysis, the 6-point rating scale was scored dichotomously, so that marks on one side indicated only agreement or disagreement.

Preliminary stages for the interpretation of the results

According to Snell, Sydell, and Lueke (1999) and Levashina and Campion (2006), individuals who have knowledge of the aspired job and the appropriate desired behaviours would be able to fake their responses better. As the applicants in our study had not received any information about the job requirements, the findings of a few faking studies that used the German edition of the PRF were considered in order to find out which scores from the PRF might show faking tendencies (with regard to the different questionnaire scales). According to a study from Stumpf and Steinhart (1981) who used the German edition of the PRF to investigate the effects of faking-good and faking-bad instructions on

a sample of soldiers and officers in training with the German Armed Forces, the following findings can be reflected upon: 1) faking-good instructions led to increased scores in the PRF scales Achievement, Affiliation, Dominance, Endurance, Exhibition, Nuturance, Order, Social Recognition, Succorance, and Understanding, and to decreased scores in the PRF scales Aggression, Harm avoidance, Impulsivity, and Play, in contrast to faking-bad instructions or standard instructions; 2) faking-bad instructions led to decreased scores in the PRF scales Achievement, Affiliation, Dominance, Endurance, Exhibition, Nurturance, Order, Social Recognition, Succorance, and Understanding, as well as to increased scores in the PRF scales Aggression, Harm avoidance, Impulsivity, and Play, in contrast to faking-good instructions or standard instructions. Altogether, faking-bad instructions led to higher differences in the scores than faking-good instructions. Rather similar effects were found in studies that used the English version of the PRF (Braun & Asta, 1969; Braun & Constantini, 1970; Hoffmann, 1968; Hoffmann & Nelson, 1971; Holden & Jackson, 1981) except for the scale Harm avoidance, where contrary results were found with respect to faking bad-instructions.

Results

To investigate the hypotheses 1, 2, and 3 a multivariate analysis of variance (MANOVA) was conducted in order to compare the means of the two experimental groups in regard to the main factor Response Format. In order to calculate the sample sizes needed to fulfil a-priori given precision requirements (type-I, type-II-risk, and a relevant effect size) we used the program CADEMO (http://www.biomath.de); however, the sample sizes were calculated according to an ANOVA. With $\alpha = .05$ and $\beta = .20$ an ANOVA is able to detect a mean difference of $\delta \geq 2/3\ \sigma$ (the standard deviation of each scale) by testing 37x2 = 74 subjects. With 84 + 184 = 268 our sample sizes were adequate. To calculate the MANOVA in order to investigate hypotheses 1 to 3, the scores of the PRF scales of the 84

subjects from Study 1 first have to be shown not to be affected by the factor Item Order. If the scores are not affected by the item order, the data from Study 1 are equivalent to the data of the 184 subjects from the current Study 2. Hence, a MANOVA with the PRF scales as dependent variables and the main factor Item Position was conducted. In the following we refer to the MANOVA with the main factor Item Position as MANOVA 1 and to the MANOVA with the main factor Response Format as MANOVA 2.

Results of the MANOVA 1 for the main factor Item Order

Box's M-Test for testing the homogeneity of the variance-covariance matrix proved to be not significant (p = .122). That is, the resulting F-values of multivariate analysis of variance can be fairly interpreted. The results of MANOVA 1 revealed no significant main effect of the factor Item Order on the PRF scales of the 84 subjects of Study 1 (p = .604; F = .886; η^2 = .160). Therefore, the data proved to be equivalent to the data of the 184 subjects from Study 2, thereby providing the necessary condition to conduct MANOVA 2.

Results of the MANOVA 2 for the main factor Response Format

The means and standard deviations of all subtests in each experimental condition are given in Table 1. Box's M-Test for testing the homogeneity of the variance-covariance matrix was significant (p = .003). To ascertain whether this significance is due to particular dependent variables (questionnaire scales) on account of the heterogeneous variances, the Levene test was calculated for each scale. The five scales Achievement, Aggression, Order, Succorance, and Infrequency were disclosed to be significant in the Levene's test (p = .009, p = .000, p = .040, p = .044, p = .022). After deleting these scales Box's M-Test proved to be non-significant (p = .276). That is, the resulting F-values of the multivariate analysis of variance can be fairly interpreted. The MANOVA for testing the main effects of the factor Response Format shows a significant effect of the Response Format (p < .001;

$F = 6.704$; $\eta^2 = .207$). The separate invariate analyses of each single scale showed significantly different means of the five scales Affiliation ($p = .021$), Endurance ($p = .010$), Harmavoidance ($p = .017$), Social Recognition ($p < .001$), and Understanding ($p < .001$) revealed significantly different means between the two experimental groups. See in Table 1 the respective means. To additionally investigate the effect of the factor Response Format on the scales Achievement, Aggression, Order, Succorance, and Infrequency, which had to be deleted from the MANOVA, two-sample *t*-tests for unequal variances (Welch tests) were applied: while significant effects resulted for the scales Aggression, Order, Succorance, and Infrequency ($p = .008$, $p = .008$, $p = .002$, $p = .027$), no significant effect occurred in regard to the scale Achievement ($p = .118$).

Insert Table 1 about here

Interpretation

The results of the MANOVA 2 support both Hypothesis 1 and Hypothesis 3, but not Hypothesis 2. The factor Response format has a significant main effect on an applicant's response behaviour. The independent analyses of the single scales show that nine scales of the PRF are affected by the factor Response format. The means of the experimental groups in each scale (see Table 1) are interpreted according to Stumpf and Steinhart's findings (1981) in order to ascertain which response format may lead to fewer faking tendencies. Hence, higher scores in the scales Affiliation, Endurance, Order, Social Recognition, Succorance, and Understanding, as well as lower scores in the scales Aggression, and Harm avoidance may show faking tendencies in the sense of socially desirable response distortion when the two experimental groups are compared. Additionally, lower scores in the scale Infrequency might be an indicator for faking tendencies (Hoffmann, 1968) with

respect to the experimental design. Applicants who had been given the 2-point rating scale showed higher scores in the scales Affiliation, Endurance, Succorance, and Understanding, as well as lower scores in the scales Aggression, and Infrequency than subjects who had been given the 6-point rating scale. These results lead to the assumption that the 2-point rating scale provokes higher faking tendencies than the 6-point rating scale which seems to lead to fewer faking tendencies. It has to be noted that lower scores in the scale Infrequency may be interpreted as cooperative behaviour from the test-takers rather than as faking good. The opposite is the case with the scales Order, Social Recognition, and Harm avoidance, where a 2-point rating scale led to less socially desirable responses than the 6-point rating scale. Applicants with a 6-point rating scale showed higher scores in the scales Order, and Social Recognition, as well as lower scores in the scale Harm avoidance.

Discussion

A study (quasi experiment) of a multivariate two-way design was conducted in order to investigate the influence of two different single-stimulus response formats on socially desirable response distortion within a personnel selection: a 6-point rating scale (with 1 = disagree totally to 6 = agree completely) versus a 2-point rating scale (with 1 = wrong and 2 = right). The aim was to find out whether the susceptibility of items with a single-stimulus response format can be decreased by using particular presentation modes, and to provide a contribution to the research that deals with possible interactions of variables that might moderate response behaviour. The significant main effect of MANOVA supports Hypothesis 1 as well as Hypothesis 3, but not Hypothesis 2. The factor Response Format affected an applicant's response behaviour in nine scales of the PRF. According to the means in the scales Affiliation, Endurance, Succorance, Understanding, and Aggression, applicants showed less socially desirable responses when they responded with a 6-point rating scale than when responding with a 2-point rating scale. These findings resemble

those of Khorramdel and Kubinger (2006) who were able to show that normative items were less fakable when applied with an analogue scale than with a 2-point rating scale. It might be more difficult for applicants to adjust their responses to an adopted faking good schema (Holden & Hibbs, 1995; Holden, Kroner, Fekken, & Popham, 1992) when a 6-point rating scale is used. The same results are disclosed with regard to the scale Infrequency, but the lower scores in this scale may be interpreted as cooperative behaviour from the test-takers rather than as faking good.

The means of the scales Order, Social Recognition, and Harm avoidance showed that the opposite effect occurred as a 2-point rating scale led to less socially desirable responses than the 6-point rating scale. Upon closer inspection of the content of the items of these three scales, it seems that these items were more transparent than items from the other PRF scales. The scale Harm avoidance includes a lot of items comprising adventures and risks that are considered as daily routines in the Federal Armed Forces or for pilots (e.g. going to a foreign country, parachuting, working with dangerous instruments or machines, or fighting forest fires). It is a similar case when it comes to the scales Order (e.g. making plans, hanging up clothes, arranging things, attaching importance to one's appearance) and Social Recognition (e.g. the importance of prestige, image or reputation, as well as acceptance). As transparent item contents, or items with contents that are of particular importance to a specific job are supposed to be more fakable (Furnham, 1986; Ramsay, Schmitt, Oswald, Kim, & Gillespie, 2006), we assume that the items of these three scales were highly vulnerable to faking tendencies, and that not even the 6-point rating scale could help to make what is socially desirable less transparent. Of course, this only explains why the 6-point rating scale was not less fakable, but not, why the 2-point rating did not fulfil our expectations. What is noticeable, apart from the item content regarding the scales Order, Social Recognition, and Harmavoidance, is that a lot of the appropriate items comprise extreme phrases like "never", "always" or "almost always", "by no means", or

"inexcusable". In combination with the transparent item contents these extreme phrases might have led to some kind of reactance when applicants had to respond with a 2-point rating scale, in the sense that some statements were refused that may have possibly been affirmed if they had been presented with more moderate answer possibilities (that is for example provided with the 6-point scale). In this respect, we assume that the decreased socially desirable responses are not to be interpreted as decreased fakability but rather as an underestimation of the actual trait loading.

In summary, the 6-point scale seems to be a better solution for use in personality questionnaires in personnel selection than the 2-point rating scale, as less socially desirable response distortion was revealed in most of the PRF scales. But this might not be true for all scales as this effect seems to be bound to the scale or item content. We assume that the kind of response format interacts with the item content and that items should be developed or used (with regard to their content) very carefully. We also assume that a 2-point rating scale not only enhances faking tendencies, but might also harm the measurement. Finally, we have to note that the fact that the 6-point scale showed less socially desirable responses in most scales does not mean that no intentional response distortion occurred.

According to the models of impression management which describe faking behaviour as an interaction of different variables (Goffin & Boyd, 2009; McFarland & Ryan, 2006; Mueller-Hanson, Heggestad, & Thornton, 2006), and the findings of Khorramdel and Kubinger (2006) on the combined effect of response format and time limit on faking tendencies, we suggest that the effects of a 6-point rating scale or an analogue scale on intentional response distortion might be accentuated by other variables, or might enhance their effects. At least, our experiment was able to show once more that the kind of response format has a moderating effect with respect to intentional response distortion.

Limitations and implications for further research

In fact we assume that the combination of the job-specific and transparent item contents with the extreme phrases might have lead to reactant response behaviour when the 2-point rating scale was applied, but this remains only an assumption and further research might find other or better explanations. Further research might also pay more attention to the effects of the combination of different questionnaire presentation types on response behaviour, as well as the interaction of such variables with the item content and item wording. The findings of the current study are of course limited by the sample, which is very unique as all participants were men, who had all applied for the same (pilot) training and who all came from the same institution (Austrian Federal Armed Forces). Therefore, the findings might not apply to other groups, such as women, or other occupational groups, so future research should investigate if our results can also be found in other samples. Furthermore, our experiment was only a quasi experiment as we were not able to assign the sample randomly to the experimental groups. Confounding variables might have occurred which we were ultimately not able to control.

Acknowledgments

We are grateful to the Department of Human Resources of the Austrian Federal Armed Forces, in particular Michael Mikas (Head of the Ambulance for Aviation & Traffic Psychology), Christian Czihak (Head of the Department for Aviation & Traffic Psychology), and Christian Langer (Head of the Department of Human Resources), for granting permission to test applicants within the local selection procedure, and for providing the PRF data with the 2-point rating scale. We also thank Alexander Uitz for his assistance with data collection of the PRF data with the 6-point rating scale, and project administration.

References

Borkenau, P., & Ostendorf, F. (1993). *NEO-Fünf-Faktoren Inventar (NEO-FFI)* [*NEO Five Factor Inventory (NEO FFI)*]. Manual, Göttingen: Hogrefe.

Bradley, K. M., & Hauenstein, N. M. A. (2006). The moderating effects of sample types as evidence of the effects of faking on personality scale correlations and factor structure. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*]*, 48*, 313-335.

Braun, J. R., & Asta, P. (1969). Changes in Personality Research Form scores (PRF, Form A) produced by faking instructions. *Journal of Clinical Psychology, 25*, 429-430.

Braun, J. R., & Constantini, A. (1970). Faking and faking detection on the Personality Research Form, AA. *Journal of Clinical Psychology, 26,* 516-518.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item format for applicant personality assessment. *Human Performance, 18*, 267-307.

Converse, P. D., Oswald, F. L., Imus, A., Hedricks, C., Roy, R., & Butera, H. (2008). Comparing personality test formats and warnings: Effects on criterion-related validity and test-taker reactions. *International Journal of Selection and Assessment, 16*, 155-169.

Dilchert, S., Ones, D. S., Viswesvaran, C., & Deller, J. (2006). Response distortion in personality measurement: born to deceive, yet capable of providing valid self-assessments? *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*]*, 48*, 209-225.

Ellingson, J. E., Sackett, P. R., & Hough, L. M. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*, 155-166).

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.

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personality and Individual Differences, 7*, 385-400.

Goffin, R. D., & Boyd, A. C. (2009). Faking and personality assessment in personnel selection: Advancing models of faking. *Canadian Psychology, 50*, 151-160.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behaviour. *Personnel Review, 36*, 341-355.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*, 9-24.

Hoffmann, H. (1968). Performance on the Personality Research Form under desirable and undesirable instructions: Personality disorders. *Psychological Reports, 23*, 507-510.

Hoffmann, H., & Nelson, J. I. (1971). Desirability responses in the Personality Research Form by a sample of alcoholics. *Psychological Reports, 29*, 559-562.

Holden, R., & Hibbs, N. (1995). Increment validity of response latencies for detecting fakers

on a personality test. *Journal of Research in Personality, 29,* 362-372.

Holden, R. R., & Jackson, D. N. (1981). Subtlety, information, and faking effects in personality assessments. *Journal of Clinical Psychology, 37*, 379-386.

Holden, R. R., Fekken, G. C., & Cotton, D. H. G. (1991). Assessing psychopathology using structured test-item response latencies. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3,* 111-118.

Holden, R. R., Kroner, D. G., Fekken, G. C., & Popham, S. M. (1992). A model of personality test item response dissimulation. *Journal of Personality and Social Psychology, 63,* 272-279.

Hossiep, R., & Paschen, M. (2003). *Das Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP)* [*Business-focused Inventory of Personality (BIP)*]. Manual, 2., vollständige überarb. Auflage, Göttingen: Hogrefe.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581-595.

Jackson Foldes, H., Ones, D. S., & Sinangil, H. K. (2006). Neither here, nor there: impression management does not predict expatriate adjustment and job performance. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*]*, 48,* 357-368.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced-choice offer a solution? *Human Performance, 13*, 371-388.

Kanning, U. P. (2003). Sieben Anmerkungen zum Problem der Selbstdarstellung in der Personalauswahl [Seven comments about social desirability in personnel selection]. *Zeitschrift für Personalpsychologie, 2*(4)*,* 193-197.

Karner, T, (2002). The volunteer effect of answering personality questionnaires. *Psychologische Beiträge* [latterly: *Psychological Test and Assessment Modeling*], *44,* 42-49.

Kersting, M. (2004). Zur Bedeutung der Validität und der soziale Akzeptanz in der Berufseignungsdiagnostik [The relevance of validity and applicants' acceptance of tests in personnel selection and training]. *Zeitschrift für Personalpsychologie, 3*, 83-86.

Khorramdel. L., & Kubinger, K. D. (2006). The effect of speediness on personality questionnaires: an experiment on applicants within a job recruiting procedure. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*]*, 48,* 378-397.

Kubinger, K.D. (2002). On faking personality inventories. *Psychologische Beiträge* [latterly:

Psychological Test and Assessment Modeling]*, 44,* 10-16.

Lammers, F., & Frankenfeld, V. (1999). Effekte gezielter Antwortstrategien bei einem Persönlichkeitsfragebogen mit „forced-choice"-Format [Effects of selective response strategies in a personality questionnaire with forced-choice format.] *Diagnostica, 45,* 65-68.

Levashina, J., & Campion, M. A. (2006). A model of faking likelihood in the employment interview. *International Journal of Selection and Assessment, 14*, 299-316.

Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment, 14,* 131-141.

Mahar, D., Cologon, J., & Duck, J. (1995). Response strategies when faking personality questionnaires in a vocational selection setting. *Personality and Individual Differences, 18*, 605-609.

Mahar, D., Coburn, B., Griffin, N., Hemeter, F., Potappel, C., Turton, M, & Mulgrew, K. (2006). Stereotyping as a response strategy when faking personality questionnaires. *Personality and Individual Differences, 40,* 1375-1386.

Marcus, B. (2003a). Das Wunder sozialer Erwünschtheit in der Personalauswahl. [The wonder of social desirability in personnel selection settings] *Zeitschrift für Personalpsychologie, 2,* 129-132.

Marcus, B. (2003b). Persönlichkeitstests in der Personalauswahl: Sind „sozial erwünschte“ Antworten wirklich nicht wünschenswert? [Personality testing in personnel selection: Is „socially desirable“ responding really undesirable?]. *Zeitschrift für Personalpsychologie, 2,* 138-148.

Martin, B. A., Bown, C.-C., & Hunt, S. T. (2002). How effective are people at faking on personality questionnaires? *Personality and Individual Differences, 32,* 247-256.

McFarland, L. A., & Ryan, A. M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology, 85*, 812-821.

McFarland, L. A., & Ryan, A. M. (2006). Toward an integrated model of applicant faking behavior. *Journal of Applied Social Psychology, 36*, 979-1016.

Mesmer-Magnus, J., Viswesvaran, C., Deshpande, S., & Joseph, J. (2006). Social desirability: the role of over-claiming, self-esteem, and emotional intelligence. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*], *48*, 336-356.

Mueller-Hanson, M., Heggestad, E. D., & Thornton, G. C. (2003). Faking and selection: Considering the use of Personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88,* 348-355.

Mueller-Hanson, R. A., Heggestad, E. D., & Thornton, G. C. III (2006). Individual differences in impression management: An exploratory of the psychological process underlying faking. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*], *48*, 288-312.

Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test or the

potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology, 65,* 131-149.

Morgeson, F. P., Campion, M. A., Diboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007a). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683-729.

Morgeson, F. P., Campion, M. A., Diboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007b). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology, 60*, 1029-1049.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995-1027.

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.

Ones, D.S., Viswesvaran, C. &, Schmidt, F.L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology, 78,* 679-703.

Ostendorf, F., & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae revidierte Fassung (NEO-PI-R)* [*NEO Personality Inventory Revised after Costa and Mc Crae (NEO PI-R)*]. Manual, Göttingen: Hogrefe.

Ramsay, L. J., Schmitt, N., Oswald, F. L., Kim, B. H., & Gillespie, M. A. (2006). The impact of situational context variables on responses to biodata and situational judgement inventory items. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*]*, 48*, 268-287.

Robie, C., Brown, D. J., & Beaty, J. C. (2007). Do people fake on personality inventories? A verbal protocol analysis. *Journal of Business and Psychology, 21*, 489-509.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review, 16*, 155-180.

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.

Seiwald, B. B. (2002). Replicability and generalizability of Kubinger's results: Some more studies on faking personality inventories. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*], 44, 17-23.

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.

Snell, A. F., Sydell, E. J., & Lueke, S. B. (1999). Towards a theory of applicant faking: Integrating studies of deception. *Human Resource Management Review, 9,* 219-242.

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.

Stumpf, H., Angleitner, A., Wieck, T., Jackson, D. N., & Beloch-till, H. (1984). *Deutsche Personality Research Form (PRF)* [*German Version of the Personality Research Form (PRF)*]. Manual, Göttingen: Hogrefe.

Stumpf, H., & Steinhart, I. (1981). *Zur Anfälligkeit der Skalenwerte der deutschen "Personality Research Form" (KA) gegenüber tendenziöser Beantwortung* [*On the susceptibility of the scales in the German „Personality Research Form „ (KA) towards tendentious response behaviour*]. Wehrpsychologische Untersuchungen, Heft 3.

Tett, R. P., & Christiansen, N. C. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007). *Personnel Psychology, 60*, 967-993.

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.

Winkelspecht, C., Lewis, P., & Thomas, A. (2006). Potential effects of faking on the NEO-PI-R: Willingness and ability to fake changes who gets hired in simulated selection decisions. *Journal of Business and Psychology, 21*, 243-259.

Wright, S. S. & Miederhoff, P. A. (1999). Selecting students with personal characteristics relevant to pharmaceutical care. *American Journal of Pharmaceutical Education, 63*, 132-138.

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.

Ziegler, M., Schmidt-Atzert, L., Bühner, M. & Krumm, S. (2007). Fakability of different measurement methods for achievement motivation: questionnaire, semi-projective, and objective. *Psychology Science* [latterly: *Psychological Test and Assessment Modeling*]*, 49*, 291-307.

Table 1:

Means and standard deviations for all dependent variables (scales) in the experimental condition Response Format (2 experimental groups)

Dependent Variable (Scale)	2-Point Rating Scale (n=184)		6-Point Rating Scale (n=84)	
	Mean	Standard Deviation	Mean	Standard Deviation
Achievement	13.516	1.406	13.149	1.918
Affiliation	14.353	2.180	13.673	2.298
Aggression	4.321	2.132	5.238	2.783
Dominance	12.342	2.832	13.036	2.745
Endurance	13.669	2.204	12.893	2.445
Exhibition	10.158	2.992	10.768	2.638
Harm avoidance	3.745	2.633	2.905	2.675
Impulsivity	4.913	2.610	4.726	2.251
Nurturance	11.717	2.306	11.649	2.426
Order	12.147	3.163	13.155	2.724
Play	9.761	3.117	9.321	2.862
Social Recognition	8.163	2.949	9.631	2.610
Succorance	7.283	2.583	6.310	2.270
Understanding	9.837	2.721	8.411	3.060
Infrequency	.457	.660	.691	.846

Running Head: **Stability of test order effects**

Test order effects: a one-hit-wonder?

Trying to replicate the findings of Khorramdel and Frebort (2010)

Lale Khorramdel and Martina Frebort

University of Vienna

Lale Khorramdel and Martina Frebort, Center of Testing and Consulting, Division of Psychological Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna.

Correspondence concerning this article may be addressed to Lale Khorramdel, University of Vienna, Liebiggasse 5, 1010 Vienna, Austria, Email: lale.khorramdel@univie.ac.at, or Martina Frebort, Email: martina.frebort@univie.ac.at

Abstract

The findings of Khorramdel and Frebort (accepted für publication 2010) about decreased "frustration tolerance" and increased "decisiveness" using objective personality tests sensu R. B. Cattell administered after cognitive ability tests (Experiment 1) were investigated once again. The same experiment was performed to investigate the effects of a varied test order of a computer based test battery on the test performance of managers. While Experiment 1 included persons in "higher management positions", the current experiment (Experiment 2) deals with incumbents of the same Austrian industrial corporation (an automotive supplier) who were in "lower positions" (shift foremen and machinery adjusters) but still had managerial responsibilities. They too attended an investigation of their professional potential which resembles a real selection situation. The results of Experiment 1 could not be replicated; no main effect of test order occurred, but there was an effect for a single test score: subjects who worked on cognitive ability tests subsequent to objective personality tests showed a significantly lower "memory" than subjects who worked on cognitive ability tests before objective personality tests. The results of the two experiments are discussed with respect to subjects' differences in cognitive ability, achievement motivation, and resilience, and with respect to differences in the test batteries.

Key words: context effects, test order, frustration tolerance, objective personality tests, cognitive ability tests, selection, assessment

Introduction

The influences of different contexts on response behaviour and test performance have received considerable attention over the past years, and are of course an iteressting and important contribution to psychological research. Context effects are mainly investigated with respect to item and task order in questionnaires (Hartig, Hölzel & Moosbrugger, 2007; Knowles, 1988; Knowles et al., 1992; Rost, & Hoberg, 1996) and cognitive ability and achievement tests (Leary & Dorans, 1985; Perlini, Lind & Zumbo, 1998). Considering that context effects occur as different orders may influence all stages of information processing underlying response behaviour (interpretation, retrieval of information, rendering a judgement, selection of a response; see Tourangeau & Rasinski, 1988), it can be concluded that this influence may pertain to different test orders as well. Particular carry-over effects, priming effects, fatigue effects, and learning effects might take place in regard to different test orders, as they involve the transfer of prior content, meaning, or behaviour and influence subsequent reactions. Regarding some of the few studies that focus on varied test sequences, it seems that varying the order of personality questionnaires has no effects on response behaviour (Eisenhauer, 2008) while cognitive ability tests affect responses on particular scales (spontaneous aggressiveness, emotional lability) of a personality questionnaire when administered first (Hambros, 2002). Increased working speed and a decreased number of correct answers towards the end of a test were found by varying test sequences of different cognitive ability tests (Eiselt, 1991); subjective self-reported motivation was made responsible while self-reported subjective fatigue had no effect. Learning effects as well as fatigue effects were found due to later positions in the test sequence by varying the sequence of tests and subtests that measured attention and memory (Földényi, Tagwerker-Neuenschwander, Giovanoli, Schallberger, & Steinhausen, 1999; Zhu & Tulsky, 2000). But results were either inconsistent or the effect sizes were

rather small. Learning effects were also found with respect to the test order of a figure test measuring field dependence-independence when the more difficult subtest was administered first (Kelleher, McRae, & Young, 1990). Khorramdel and Frebort (accepted für publication 2010) found that administering cognitive ability tests before objective personality tests increased the test score "decisiveness" and decreased the test score "frustration tolerance" of the objective personality test *Work Styles* ("Arbeitshaltungen"; Kubinger, & Ebenhöh, 2007), while there was no reverse effect. Equal effects were found for the test score "decisiveness" of the *Work Styles* when this test was administered after a personality questionnaire (Baldinger, 2006) in contrast to the administration before the questionnaire where no effects were found. A basical model for the explanation of context effects is offerd by Schwarz, Hippler and Noelle-Neumann (1992), who argue that order effects (in dimensional sets of response alternatives) depend on a complex interaction of serial position, presentation mode, item attributes (e.g. plausibility, complexity, and extremity of the wording), and respondents' ability and motivation, as described by the elaboration likelihood model (Petty & Cacioppo, 1986). Considering the factor of respondets' ability, it was found that response-order effects were greater among respondents with less cognitive sophistication, which led to the suggestion that response-order effects are the result of inadequate memory search and superficial response alternative evaluation (Krosnick, 1992). It was also shown that individuals differ in cognitive evaluation processes depending on their cognitive complexity (O'Keefe, Delia, & O'Keefe, 1977). Subjects with higher cognitive complexity were shown to organize information at a higher level than individuals with lower cognitive complexity, who had less flexible schemas (O'Keefe, Delia, & O'Keefe, 1977) but were better able to integrate conflicting stimulus information into their concepts. Nevertheless, apart from order effects, these subjects (with high and low cognitive complexity) did not differ in their impression evaluations (toward a stimulus person)

despite their different cognitive processes. Finally, the appearance and the extent of context effects might additionally, or sometimes even primarily, be influenced by subjects' motivation (Hippler, & Schwarz, 1987). Most of the studies focusing on order effects did not consider these variables and hence did not provide a satisfying explanation of order effects.

Aims of the current experiment

An experiment investigating the effects of varied test orders with respect to so-called objective personality tests is presented. The use of different test orders within test batteries is a common practice because of different reasons, such as organisational reasons (for example not all test are available on all computers), to avoid cheating, or to avoid a decrease of motivation if some test-takers realise that others are perform faster. But this practice is not well explored or proven to be without consequences for the test results. The aim of the current experiment is to investigate if the findings of Khorramdel and Frebort (accepted für publication 2010) can be replicated. A further aim was to find out if there are any significant effects of changed test order due to subjects' differences in cognitive ability and achievement motivation. Therefore, the sample of the current experiment (Experiment 2) consists of persons with a lower level of education compared with the sample of the experiment of Khorramdel and Frebort (accepted für publication 2010) (Experiment 1), assuming that this lower ecucational level is, in general, accompanied by lower cognitive abilities or lower achievement motivation.

Materials and Methods

Sample

As in Experiment 1, the sample of the current experiment (Experiment 2) consists of 64 incumbents of an Austrian industrial corporation (an automotive supplier) with managerial responsibilities – this time in "lower positions" within the corporation hierarchy (shift foremen and machinery adjusters), in contrast to the "higher management positions" of the subjects in Experiment 1 (business managers, department chiefs, and team leaders). Again, the subjects attended an investigation of their professional potential (job-related cognitive ability and personality dimensions) to find out if they are suited for their position, if they should be given a position without managerial responsibility, or if they have the potential to obtain a higher position within the corporation. In this respect, the testing situation resembles a personnel selection situation as degradation to a position without managerial responsibility was within the bounds of possibility. None of the participants were female; the age of the participants varied between 22 to 56 years.

Measures

Similar to Experiment 1, a requirement analysis was conducted to be able to identify the requirements the participants had to meet in their everyday work. The resulting requirements profile (which comprised particular cognitive abilities, aspects of personality, and management styles) differs only marginally from the one in Experiment 1. According to this profile an almost equal test battery was arranged. In contrast to Experiment 1, the AMT (Adaptive Test of Matrices – German edition; Hornke, Etzel, & Rettig, 2007) was replaced by the SPM (Standard Progressive Matrices – German edition; Raven, Raven, & Court, 2008), and the two subtests verbal and numerical intelligence of the IST 2000 R

(Intelligence Structure-Test – German edition; Liepmann, Beauducel, Brocke, & Amthauer, 2007) were supplemented with the subtest memory.

Cognitive ability tests

Standard Progressive Matrices (SPM) – German edition (Raven, Raven, & Court, 2008). The SPM is a reasoning test where items are presented conventionally: ascending from easy to hard (in contrast to the adaptive presented items of the AMT). The SPM differentiate better between persons with lower abilities as it comprises more items that are more likely to be solved than the items of the AMT.

Intelligence Structure Test (IST 2000 R) – German edition (Liepmann, Beauducel, Brocke, & Amthauer, 2007). The IST 2000 R is an intelligence test battery consisting of 11 subtests that measure verbal (3 subtests), numerical (3 subtests), and figural (3 subtests) intelligence, as well as memory (2 subtests). For the current experiment, only the subtests measuring verbal intelligence, numerical intelligence, and memory were selected.

Objective Personality Tests

Objective personality tests sensu R. B. Cattell (e.g. 1958) are experiment-based assessments of behaviour which assess a personality construct by observing the subject's behaviour when working on a performance or ability task, while the observation and registration of the behaviour is done via computer (Kubinger, 2009).

Work Styles – German edition (Kubinger, & Ebenhöh, 2007). The test-battery consists of three subtests. Subtest 1 ("comparing area sizes") measures decisiveness, exactitude, and reflexivity, Subtest 2 ("coding symbols") measures proficiency level, aspiration level, target discrepancy, and frustration tolerance, and Subtest 3 ("distinguishing figures") measures achievement motivation.

Resilience-Assessment: computer based Objective Personality Test Battery (BAcO-D) – German edition (Ortner, Kubinger, Schrott, Radinger, & Litzenberger, 2006). BAcO-D is a test battery that measures different kinds of job-related resilience, and consists of six subtests. For the current experiment two of the six subtests were selected: the subtest "task collision" which measures one's resilience given multiple simultaneous tasks, and the subtest "crisscrossed plans" which measures one's resilience given thwarted plans.

ILICA – a simulation test to assess decisive behaviour – German edition (Möseneder, & Ebenhöh, 1996). ILICA is a computer simulation in the German language which measures self-management abilities. A leisure day is simulated for 30 minutes, during which time arrangements for a nearing holiday have to be made.

Hypotheses

Primary hypotheses

As in Experiment 1, we expect that test performance varies depending on different test orders. We therefore formulate the following two hypotheses:

Hypothesis 1: The prior work on objective personality tests influences the subsequent performance in cognitive ability tests.

Hypothesis 2: The prior work on cognitive ability tests influences the subsequent performance in objective personality tests.

Secondary hypotheses

Corresponding to the model of Petty and Cacioppo (1986), as well as that of Schwarz, Hippler and Noelle-Neumann (1992), we hypothesise that effects of test order might occur depending on differences in subjects' cognitive ability and motivation, if the findings of Khorramdel and Frebort (accepted für publication 2010) cannot be replicated.

Design

The same design was tested as in Experiment 1. Participants were randomly assigned to two experimental groups where the sequence of cognitive ability tests and objective personality tests was varied within a computer based test battery. Group 1 completed objective personality tests first and cognitive ability tests afterwards (Test Order O: *Work Styles* – BAcO – ILICA – IST 2000 R - SPM), Group 2 completed cognitive ability tests prior to the objective personality tests (Test Order C: IST 2000 R – SPM – *Work Styles* – BacO - ILICA). The cognitive ability tests were presented in a fixed order (SPM – IST 2000 R) as were the objective personality tests (*Work Styles* – BAcO – ILICA).

Insert Figure 1 about here

Results

Primary Hypotheses

The main factor test order was investigated by conducting a multivariate analysis of variance (MANOVA). A MANOVA is able to detect a mean difference of $\delta \geq 3/4\ \sigma$ (the standard deviation of the test scores), with $\alpha = .05$ and $\beta = .20$, by testing 29x2 = 58 subjects. With 32x2 = 64, our sample size was designed adequately.

Results of the MANOVA

The means and standard deviations of all subtests in the respective experimental condition are given in Table 1. The Box's M-Test for testing the homogeneity of the variance-covariance matrix regarding all test scores was not significant ($p = .250$), indicating that the resulting *F*-values of MANOVA can fairly be interpreted. The MANOVA showed

no significant effect of test order ($p = .351$; $F = 1.138$; $\eta^2 = .396$), a result echoed in the univariate results for all scales except the score "memory" of the IST 2000 R. For this score subjects of Group 1 showed lower scores than subjects of Group 2 ($p = .041$; see the respective means in Table 1). Because the test scores "verbal intelligence" of the IST 2000 R as well as "efficiency of subsidiary task –task collision", "flexibility", and "distractibility" of ILICA revealed significant results in the Levene's test ($p = .010$, $p = .014$, $p = .002$, $p = .049$), Welch tests were applied to investigate possible effects of test order. No significant effects were revealed ($p = .318$, $p = .870$, $p = .705$, $p = .696$).

Insert Table 1 about here

Interpretation

The results of the MANOVA do not provide support for Hypothesis 2 or for Hypothesis 1. The factor test order had no main effect on subjects' test performance: previously presented cognitive ability tests did not influence test scores of subsequently presented objective personality tests, nor was the reverse true. Apart from this, independent analysis of the single test scores showed that test order influenced only the test score "memory"; subjects who worked first on objective personality tests and subsequently on cognitive ability tests showed lower memory scores.

Sample differences

To test our above mentioned assuption that the lower level of education of Experiment 2 might be accompanied by lower cognitive abilities or lower achievement motivation compared to Experiment 1 as well as the secondary hypotheses, we used Welch tests to compare the two samples in regard to those test scores that were not significantly af-

fected by the varied test order (with respect to the significant findings in Experiment 1). It was revealed that the two samples differ significantly in the scores "verbal intelligence" ($p < .001$), "numerical intelligence" ($p < .001$), "aspiration level" ($p = .032$), "target discrepancy" ($p = .016$), and "quantity – crisscrossed plans" ($p = .004$). The two samples differ significantly in the score "reasoning" ($p < .001$) as well. But as noted in the description of the measures (see above), different measures (AMT and SPM) were used in Study 1 and Study 2 to test reasoning. Therefore, this difference must be interpreted very carefully and with reservation. According to the mean scores, the sample of Experiment 1 showed higher cognitive abilities (verbal intelligence, numerical intelligence), had higher aspiration levels, and lower target discrepancies in comparison with the sample of Experiment 2. The lower target discrepancy means that subjects of Experiment 1 were better able to estimate their achievement. But they also showed lower scores in "quantity – crisscrossed plans" than subjects of Experiment 2, which means that fewer subjects of Experiment 1 reached the target in the labyrinth task, showing less resilience when their plans were thwarted. The means and standard deviations for all scores that significantly differ between the two samples are given in Table 2.

Insert Table 2 about here

Discussion

The findings of Experiment 1 (Khorramdel, & Frebort, accepted für publication 2010) could not be replicated in the current experiment (Experiment 2). No significant main effect of the factor test order was revealed in the MANOVA, providing no evidence for Hypothesis 1 or Hypothesis 2. Working on cognitive ability tests did not lead to higher decisiveness or lower frustration tolerance in the *Work Styles* as it did in Experiment 1. Similar to the findings of Zhu and Tulsky (2000), test order influenced only the test score "memory", which could be interpreted as a slight fatigue effect, and which might be ascribed more to the test length than to the previously presented objective personality tests. Subjects who worked on cognitive ability tests after objective personality tests showed lower scores than subjects who worked on cognitive ability tests first. The differences we found between the two samples regarding their cognitive abilities, aspiration levels, and resilience (in regard to crisscrossed plans) provide evidence for our secondary hypothesis, which assumes that effects of test order might occur depending on differences in subjects' cognitive ability and motivation. Subjects in Experiment 1 showed higher cognitive abilities and higher aspiration levels but were less resilient when their plans were impeded than subjects in Experiment 2. According to Heckhausen and Heckhausen (2006), persons with high achievement tend to have medium to high aspiration levels and realistic self-estimations in regard to their achievements, something that is shown in Experiment 1 (see "Sample differences" above). It can be assumed that the subjects of Experiment 1 are less resilient when their plans do not work out because they have a high aspiration level and success is important to them, while subjects in Experiment 2 might be more resilient because success is less important to them (which does not mean that it is not important to them at all). Additionally, the frustration effect of the *Working Styles* is provoked through a social comparison in the form of faked feedback (within a simple symbol coding task)

that others have achieved higher scores. Success and high achievements, and therefore social comparisons through which both are reflected, might have been more important to subjects of Experiment 1. The combination of this importance with the kind of frustration provoked could explain the significantly lower frustration tolerance that only occurred in Experiment 1, as we theorized in Khorramdel and Frebort (accepted für publication 2010). Regarding the effects due to the score decisiveness, that occurred only in Experiment 1 as well, it seems interesting, that the sample of Baldinger (2006), whose findings about the score decisiveness of the *Work Styles* we were able to replicate in Experiment 1, was also characterised by a higher educational level.

Comparing Experiment 1 with Experiment 2 with respect to the applied measures could provide another possible explanation for our different findings. The almost identical test batteries differ in one point: different matrices tests were applied. This could be one important explanation for the different findings in the two experiments. The sample in Experiment 1 received the adaptive AMT, where the item difficulty is adjusted depending on the subject's ability, and the sample in Experiment 2 received the SPM, where items are presented conventionally, ascending from easy to difficult. Through the AMT, subjects had to deal with more difficult items than the subjects who worked on the SPM. Thus, the adaptively presented items might have caused higher exertion and therefore intensified the provoked frustration effect of the *Work Styles*. That adaptive testing has an effect on subjects' motivation is shown by Frey, Hartig and Moosbrugger (2009). They revealed that the test-taking motivation (measured with a questionnaire) of subjects who worked on an adaptive version of a concentration test was significantly lower than the test-taking motivation of subjects who worked on a non-adaptive version; this can be ascribed to a perceived lower probability of success.

Altogether, the following conclusions can be drawn: 1) Test order might only affect particular kind of subjects or samples as it affected particular scores ("decisiveness", "frustration tolerance") only in a sample with higher cognitive abilities and status that additionally showed higher aspiration levels and lower resilience. 2) This effect of test order occurs only in subtests with very simple tasks (like *Work Styles*), while more complex tasks (like AMT, SPM, IST 2000 R) are not affected. 3) There might be a different influence of the two matrices tests SPM and AMT on the score frustration tolerance of the *Work Styles* because the AMT is an adaptive test and therefore tends to consist of more difficult items than the SPM. The adaptive items might cause higher exertion and therefore intensify the frustrating situation in the *Work Styles*.

Practical implications

Our findings were inconsistent as the results of Experiment 1 could not be replicated in Experiment 2. Depending on the kind of sample and the kind of measures, order effects may or may not occur. When in doubt, the effect of different test sequences should not be neglected. Particularly in psychological assessments when test scores of applicants are compared in order to identify the most qualified ones – the test order of test batteries comprising objective personality tests or comparable computer simulations should be held constant for all participants. On the other hand, if particular effects (such as the extent of frustration tolerance) are intended, they might be controllable if more were known about test order effects.

Limitations and implications for future research

Our results showed that the revealed context effects of Experiment 1 were not replicable in Experiment 2. This may be due to the different matrices tests used (an adaptive

versus a conventional one) or to the different composition of the samples, as subjects in Experiment 1 were more highly educated, had higher cognitive abilities, higher aspiration levels, and held higher job position than subjects in Experiment 2. As context effects seem to depend on the kind of sample, the effects revealed in Experiment 1 might not be generalized to other ability tests or personality measures. Further research should focus systematically on the composition of the samples and the content of the items (or tests) used. A comprehension of the possible interactions between sample, task type, content of the measurement, and the modality of specific simulations might reveal helpful explanations for context effects. Our assumption of the effects of the different matrices tests (AMT, SPM) should be investigated separately from other moderating variables (such as differences in motivation or cognitive ability). Furthermore, it would also be interesting to investigate possible order effects by varying the order of different objective personality tests; their order was held constant in the current experiment. The current study has a limitation that future research should address. The sample is very special as most participants were men (only four of the participants were female), and the participants were all managers in “higher management positions” of a particular industrial corporation (an automotive supplier). Therefore, the findings might not apply to other groups, such as women, other occupational groups, or other positions within similar professions.

References

Baldinger, D. (2006). *Der Einfluss sozial erwünschten Verhaltens auf das Ergebnis Objektiver Persönlichkeitstests [The influence of social desirable behaviour on the results of objective personality questionnaires]*. Unpublished diploma thesis, University of Vienna.

Cattell, R. B. (1958). What is "objective" in "objective" personality tests?. *Journal of Counseling Psychology, 5*, 285-289.

Eiselt, W. (1991). *Der Einfluss subjektiver Daten und Testreihenfolge auf die Leistung in einer computergestützten Testbatterie [The influence of subjective data and test order on the achievement within a computer based test battery]*. Investigation of the psychological office of the German armed forces (Navy Medical Institute of the Marine), Kiel, 5-54.

Eisenhauer, E. (2008). *Effekte bei der Veränderung der Itempositionen anhand des Anstrengungsvermeidungstest und B5PO*. Unpublished thesis, University of Vienna.

Földényi, M., Tagwerker-Neuenschwander, F., Giovanoli, A. Schallberger, U., & Steinhausen, H.-C. (1999). Die Aufmerksamkeitsleistungen von 6-10-jährigen Kindern in der TAP [Attentional performance of 6-10-year-old children on the TAP]. *Zeitschrift für Neuropsychologie, 10,* 87-102.

Frey, A., Hartig, J., & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationstleistungs-Tests [Effects of adaptive testing on test-taking motivation with the example of the Frankfurt Adaptive Concentration Test]. *Diagnostica, 55*, 20-28.

Hambros, K. (2002). On reasonableness of personality inventories with dichotomous item response format. *Psychologische Beiträge* [latterly: *Psychological Test and Assessment Modeling*]*, 44,* 126-135.

Hartig, J., Hölzel, B., & Moosbrugger, H. (2007). A confirmatory analysis of item reliability trends (CAIRT): Differentiating true score and error variance in the analysis of item context effects. *Multivariate Behavioral Research, 42,* 157-183.

Heckhausen, J., & Heckhausen, H. (2006). *Motivation und Handeln [Motivation and Action]*. 3. edition, Heidelberg: Springer.

Hippler, H.-J., & Schwarz, N. (1987). Response effects in surveys. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social information processing and survey methology,* (pp. 102-112). New York: Springer.

Hornke, L. F., Etzel, S., & Rettig, K. (2007). *Adaptiver Matrizentest (AMT) [Adaptive Matrices Test]*. Version 26.00, Mödling: Schuhfried.

Kelleher, W. E., McRae, L. S. E., & Young, J. D. (1990). The group embedded figure test: The learning effect reexamined. *Perceptual and Motor Skills, 70,* 1233-1234.

Khorramdel, L., & Frebort, M. (accepted for publication 2010). Context effects on test performance: What about test order? *European Journal of Psychological Assessment*.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55*, 312-320.

Knowles, E. S., Coker, M. C., Cook, D. A., Diercks, S. R., Irwin, M. E., Lundeen, E. J., Neville, J. W., & Sibicky, M. E. (1992). Order effects within personality measures. In N. Schwarz & S. Sudman. *Context effects in Social and Psychological Research,* (pp. 221-236). New York: Springer.

Krosnick, J. A. (1992). The impact of cognitive sophistication and attitude importance on response-order and question-order effects. In N. Schwarz & S. Sudman. *Context effects in Social and Psychological Research,* (pp. 203-218). New York: Springer.

Kubinger, K. D., & Ebenhöh, J. (2007). *Arbeitshaltungen (AHA) [Work Styles]*. Version 27.00, Mödling: Schuhfried.

Kubinger, K. D. (2009). The technique of objective personality-tests sensu R. B. Cattell nowadays: The Viennese pool of computerized tests aimed at experiment-based assessment of behaviour. *Acta Psychologica Sinica*, *41*, 1024-1036.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55,* 387-413.

Liepmann, D., Beauducel, A., Brocke. B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R (IST 2000 R) [Intelligence Structure Test]*. 2. revised edition, Göttingen: Hogrefe.

Möseneder, D., & Ebenhöh, J. (1996). *Ein Simulationstest zur Erfassung des Entscheidungsverhaltens (ILICA) [A simulation test for the measure of decision behaviour]*. Frankfurt: Swets.

O'Keefe, B. J., Delia, J. G., & O´Keefe, D. J. (1977). Construct individuality, cognitive complexity, and the formation and remembering of interpersonal impressions. *Social Behavior and Personality, 5,* 229-240.

Ortner, T. M., Kubinger, K. D., Schrott, A., Radinger, R., & Litzenberger, M. (2006). *Belastbarkeits-Assessment: computerisierte Objektive Persönlichkeits-Testbatterie – Deutsch (BAcO-D) [Resilience-Assessment: computer based Objective Personality Test Battery – German version]*. Frankfurt/M.: Harcourt Test Services.

Perlini, A. H., Lind, D. L., & Zumbo, B. D. (1998). Context effects on examinations: The effects of time, item order and item difficulty. *Canadian Psychology/Psychologie canadienne, 39*, 299-307.

Petty, R. E., & Cacioppo J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. New York: Springer-Verlag.

Raven, J., Raven, J. C., & Court, J. H. (2008). *Standard Progressive Matrices (SPM)*. Version 31.04, Mödling: Schuhfried.

Rost, D. H., & Hoberg, K. (1996). *Itempositionsveränderungen in Persönlichkeitsfragebogen: Methodischer Kunstfehler oder tolerierbare Praxis? [Changing the position of items in personality questionnaires: Methodological malpractice or tolerable practice?]*. Reports of the Department of Psychology, Vol. 114, Philipps-University Marburg.

Schneider, M., & Stern, E. (2010). The developmental relations between conceptual and procedural knowledge: A multimethod approach. *Developmental Psychology, 46,* 178-192.

Schwarz, N., Hippler, H.-J., & Noelle-Neumann, E. (1992). Cognitive model of response-order effects. In N. Schwarz & S. Sudman. *Context effects in Social and Psychological Research, (*pp. 187-201). New York: Springer.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103,* 299-314.

Zhu, J., & Tulsky, D. S. (2000). Co-norming the WAIS-III and WMS-III: Is there a test-order effect on IQ and memory scores? *The Clinical Neuropsychologist, 14*, 461-467.

Table 1:
Means and standard deviations for all subtests in each experimental condition (2 experimental groups) with regard to the factor "test order"

Dependent variable (test score)	Test order O		Test order C	
	Mean	Standard deviation	Mean	Standard deviation
Reasoning – AMT	-.119	.770	.001	.598
Verbal intelligence – IST 2000 R	-1.135	1.501	-.798	.948
Numerical intelligence – IST 2000 R	-.398	.915	-.201	.772
Memory – IST 2000 R	-.524	.852	-.113	.714
Reflexivity – *Work Styles*	.104	.983	.155	.852
Exactitude – *Work Styles*	.393	1.112	.416	.993
Decisiveness – *Work Styles*	-.380	.685	-.239	.742
Proficiency level – *Work Styles*	.358	.745	.395	.617
Aspiration level – *Work Styles*	-.405	.707	-.146	.846
Frustration tolerance – *Work Styles*	.231	.807	.131	.700
Target discrepancy – *Work Styles*	.091	1.148	.092	.828
Main task – BAcO-D task collision	-.037	1.004	-.134	1.087
Efficiency of subsidiary task – BAcO-D task collision	-.380	.918	-.418	.640
Quantity of subsidiary task – BAcO-D task collision	-.330	1.145	-.039	1.049
Perseverance – BAcO-D task collision	-.072	1.018	-.472	1.207
Balance – BAcO-D task collision	.805	1.909	.233	1.642
Quantity – BAcO-D crisscrossed plans	3.270	1.419	3.560	.906
Speed – BAcO-D crisscrossed plans	-.268	.978	-.265	.802
Abidance – BAcO-D crisscrossed plans	.885	.672	.895	.556
Stray – BAcO-D crisscrossed plans	.422	1.060	.366	.890
Flexibility – ILICA	-.221	1.108	-.315	.573
Target orientation – ILICA	.523	1.026	.670	.773
Distractibility – ILICA	.136	1.088	.060	.714

Table 2:
Means and standard deviations for all scores that (according to the Welch test) differ significantly between the two samples of Experiment 1 and Experiment 2

Dependent variable (test score)	Sample – Experiment 1		Sample – Experiment 2	
	Mean	Standard deviation	Mean	Standard deviation
Verbal intelligence – IST 2000 R	.219	.925	-.990	1.246
Numerical intelligence – IST 2000 R	.525	.964	-.344	.851
Aspiration level – *Work Styles*	-.012	.822	-.312	.787
Target discrepancy – *Work Styles*	-.333	1.074	.099	.977
Quantity – BAcO-D crisscrossed plans	2.449	2.222	3.362	1.278

7. Curriculum Vitae

Lale Khorramdel, M.Sc. (Mag. rer. nat.)

Personal Data	* 25.12.1978 in Vienna Citizenship: Austrian

Education

since 10/2004	PhD studies in psychology (Dr. rer. nat.), University of Vienna
2005	Graduate training in work and organisational psychology
07/2004 – 09/2005	Graduate training in clinical and health psychology
2005	Licence A, DIN 33430 (job related proficiency assessment), Academy of German Psychologists (DPA), Association of German Professional Psychologists (BDP)
10/1998 – 05/2004	Studies of psychology (Mag. rer. nat., Diplomstudium), University of Vienna
10/1997 – 06/1998	Studies of graphic design and art classes, education centre of arts, Vienna

Work experience

since 11/2007	Assistant Director of the *Centre of Testing and Consulting*, *Division of Psychological Assessment and Applied Psychometrics* (Director: Univ.-Prof. Dr. Klaus D. Kubinger), Faculty of Psychology, University of Vienna
10/2004 – 10/2007	Research Assistant and Assistant Project Coordinator of the *Centre of Testing and Consulting*, *Division of Psychological Assessment and Applied Psychometrics* (Director: Univ.-Prof. Dr. Klaus D. Kubinger), Faculty of Psychology, University of Vienna
since 03/2006	Lecturer in *Psychological Assessment and Applied Psychometrics*, Faculty of Psychology, University of Vienna
10/2005 – 02/2006	Teaching Assistant, *Division of Psychological Assessment and Applied Psychometrics* (Director: Univ.-Prof. Dr. Klaus D. Kubinger), Faculty of Psychology, University of Vienna

8. Publications

Journal Articles

Published

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. Psychology Science Quarterly, 50, 391-402.

Khorramdel, L., & Frebort, M. (accepted for publication). Context effects on test performance: What about test order? *European Journal of Psychological Assessment*.

Khorramdel, L., & Kubinger, K. D. (2006). The Effect of Speediness on Personality Questionnaires: An Experiment on Applicants Within a Job Recruiting Procedure. *Psychology Science, 48*, 378-397.

Khorramdel, L. & Kubinger, K. D. (2008). Psychologische Instrumente in der Personalauswahl: Chancen der experimentellen Verhaltensdiagnostik [Psychological measurements in personnel selection: Chances of experiment-based assessments of behaviour]. *Personalmanager – Zeitschrift für Human Resources, 2008/4*, 31-33.

Kubinger, K. D., Frebort, M., Holocher-Ertl, S., Khorramdel, L., Sonnleitner, P., Weitensfelder, L., Hohensinn, C., & Reif, M. (2007). Large-Scale Assessment zu den Bildungsstandards in Österreich: Testkonzept, Testdurchführung und Ergebnisverwertung [Large-scale assessment at the Austrian educational standards: test concept, test implementation and application of the results]. *Erziehung und Unterricht, 7-8*, 588-599.

Kubinger, K. D., Frebort, M., Khorramdel, L., Weitensfelder, L., Sonnleitner, P., Hohensinn, C., Reif, M., Gruber K., & Holocher-Ertl, S. (2008) Large Scale Assessment at the Austrian Educational Standards: a Review. Testing International, 19, 15-16.

Submitted

Khorramdel, L., Kubinger, K. D., & Uitz, A. (submitted). Questionnaire length and impression management: Do applicants just forget to fake? *International Journal of Selection and Assessment*.

Khorramdel, L., & Kubinger, K. D. (submitted). The influence of different rating scales on impression management. Should we give um on rating scales? *Human Performance.*

Khorramdel, L., & Frebort, M. (submitted). Test order effects: a one-hit-wonder? Trying to replicate the findings of Khorramdel and Frebort (2010). *European Journal of Psychological Assessment*.

Kubinger, K. D., Hohensinn, C., Hofer, S., Khorramdel, L., Frebort, M., Holocher-Ertl, S., Reif, M., & Sonnleitner, P. (submitted), Designing the test booklets for Rasch model calibrations in a large scale assessment with reference to numerous moderator variables and multiple ability dimensions. *Educational Research and Evaluation*.

Readers

Khorramdel, L. (in press). Potenzialanalyse zur Führungskräfteentwicklung – Herr M., 25 Jahre [Analysis of potential regarding development of managers – Mr. M., 25 years]. In K. D. Kubinger &M. Ortner. *Psychologische Diagnostik in Fallbeispielen*. Göttingen: Hogrefe.

Frebort, M., & Khorramdel, L. (in press). Auswahl von Tierpflegerschülern – der Jahrgang 2009/10 [Selection of animal care students 2009/10]. In K. D. Kubinger &M. Ortner. *Psychologische Diagnostik in Fallbeispielen*. Göttingen: Hogrefe.

Congress Contributions

Presentations

Frebort, M., Khorramdel, L., Kubinger, K. D., & Maurer, M. (2008). *An instruction guide for non-psychologists to administrate a test: Experiences of a large scale assessment using teachers.* 29th International Congress of Psychology, Berlin, Germany, 20th-25th July 2008 (Symposium).

Hohensinn, C., Holocher-Ertl, S., Kubinger, K. D., Reif, M., Khorramdel, L. & Frebort, M. (2007). *Large-Scale Assessment: Grundlagenforschung zu Item-Reihenfolgeeffekten* [*Lare-scale assessment: Basic research on effects of item position*]. 9. Arbeitstagung der Fachgruppe für Differentielle Psychologie, Persönlichkeitspsychologie und Psychologische Diagnostik**,** Wien, Österreich, 24.-26. September 2007.

Hohensinn, C., Kubinger, K. D., Holocher-Ertl, S., Reif, M., Khorramdel, L., & Frebort, M. (2007). *„Potenzialanalyse" Rasch-Modell-konformer Itempools mit Hilfe des Difficulty-plus-Guessing PL-Modells nach Kubinger & Draxler* [*„Potential analysis" of Rasch model fitting item pools by means of the difficulty-plus-guessing PL-Model of Kubinger & Draxler*]. 4. Klangenfurter Statistiktag: Statistik in der psychologischen Forschung, Klangenfurt. 19.-20. Oktober 2007.

Hohensinn, C., Kubinger, K. D., Holocher-Ertl, S., Reif, M., Khorramdel, L., Sonnleitner, P., Gruber, K. & Frebort, M. (2008). *Artifizielle Testwerte bei inadäquater Berücksichtigung von missing values* [*Artificial test scores due to inadequate consideration of missing values*]. 8. Wissenschaftliche Tagung der Österreichischen Gesellschaft für Psychologie, Linz, April 2008.

Hohensinn, C., Kubinger, K. D., Holocher-Ertl, S., Reif, M., Khorramdel, L., Sonnleitner, P., Gruber, K., & Frebort, M.: (2008). *Measuring the effect of multiple choice response format by some IRT models.* XXIXth International Congress of Psychology, Berlin, Germany, 20th-25th July 2008.

Khorramdel, L. & Kubinger, K. D. (2006). *Speed als Mittel gegen die Verfälschbarkeit von Persönlichkeitsfragebogen? Ein Experiment* [*Speed appliance against the fakability of personality questionnaires) An experiment*]. 7. Wissenschaftliche Tagung der Österreichischen Gesellschaft für Psychologie, Klagenfurt, 28.-30. April 2006.

Khorramdel, L. & Kubinger, K. D. (2007). *Speed als Mittel gegen die Verfälschbarkeit von Persönlichkeitsfragebogen? Ein Experiment* [*Speed appliance against the fakability of personality questionnaires) An experiment*]. 9. Arbeitstagung der Fachgruppe für Differentielle Psychologie, Persönlichkeitspsychologie und Psychologische Diagnostik**,** Wien, Österreich, 24.-26. September 2007.

Khorramdel, L., & Kubinger, K. D. (2009). *Effects of the variation of test order within a computer test battery for the evaluation of managers´ aptitude*. 10th European Conference on Psychological Assessment, Ghent, Belgium, 16th-19th September 2009.

Khorramdel, L. & Kubinger, K. D. (2009). *Reihenfolgeeffekte innerhalb einer Computertestbatterie zur Eignungsbeurteilung von Führungskräften* [*Effects of the variation of test order within a computer test battery for the evaluation of managers´ aptitude*]. 10. Arbeitstagung der Fachgruppe für Differentielle Psychologie, Persönlichkeitspsychologie und Psychologische Diagnostik**,** Landau, 28.-30. September 2009.

Khorramdel, L. & Kubinger, K. D. (2010, accepted). *Die Verfälschbarkeit als möglicher Kontexteffekt: wie wirken sich eine übermäßige Testlänge und unterschiedliche Antwortformate aus* [*Fakability as possible context effect: what effects do an extensive test length and different response formats have*]? 47. Kongress der Deutschen Gesellschaft für Psychologie, Bremen, 26.-30. September 2010.

Khorramdel, L., Kubinger, K. D. & Punter, J.F. (2004). *Und noch ein Experiment zur Verfälschbarkeit von Persönlichkeitsfragebögen* [*Another experiment on the fakability of personality questionnaires*]. 6. Tagung der Österreichischen Gesellschaft für Psychologie, Innsbruck, 26.-28. Februar 2004.

Khorramdel, L., Kubinger, K. D., & Punter, J.F. (2005). *The effect of speeded administration on faking personality questionnaires in selection situations*. 9th European Congress of Psychology, Granada, Spain, 3rd – 8th July, 2005.

Khorramdel, L., Kubinger, K. D., & Punter, J.F. (2005). *The effect of speeded administration on faking personality questionnaires in selection situations*. 8th European Conference of Psychological Assessment, Budapest, Hungary, 31st August – 4th September, 2005.

Khorramdel, L., Punter, J.F. & Kubinger, K. D. (2004). *Der Einfluss der Vorgabebedingung „speed" auf die Verfälschbarkeit von Persönlichkeitsfragebogen in Auswahlsituationen* [*The influence of speediness on the fakability of personality questionnaires in personnel selection*]. 44. Kongress der Deutschen Gesellschaft für Psychologie, Göttingen, 26.-30. September 2004.

Poster

Khorramdel, L., Maurer, M., & Kubinger, K. D. (2008). *A requirement analysis of study specific demands – What requirements of ability and personality do students need to be successful?* 29th International Congress of Psychology, Berlin, Germany, 20th-25th July 2008.