Predictors of Performance: The Impact of Source, Domain-Specificity, and Structure

Dissertation

zur Erlangung des akademischen Grades

Dr. rer. nat. im Fach Psychologie

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät II

der Humboldt-Universität zu Berlin

von Dipl. Psych. Erik Danay

Präsident der Humboldt-Universität zu Berlin Prof. Dr. Jan-HendrikOlbertz

Dekan der Mathematisch-NaturwissenschaftlichenFakultät IIProf.Dr. Peter Frensch

Gutachter/in: 1. Prof. Dr. Matthias Ziegler

2. Prof. Dr. Markus Bühner

3. Prof. Dr. Thomas Götz

Datum der Einreichung: 22.12.2010

Datum der Promotion: 21.2.2011

Eidesstattliche Erklärungen

Hiermit versichere ich des Eides statt, dass

a) ich die vorliegende Dissertation mit demTitel "Predictors of Performance: The Impact of Source, Domain-Specificity, and Structure" selbständig und ohne unerlaubte Hilfe angefertigt habe.
b) essich um die Ersteinreichung der vorliegendenArbeitals Dissertation handelt.
c) ich die Promotionsordnung der Humboldt-Universität zu Berlin zur Kenntnis genommenhabe.

Berlin, 20.12.2010 Erik Danay

Contents

Zusammenfassung

Diese publikationsorientierte Dissertation umfasst drei Arbeiten zum Thema der Prädiktion von Leistung. In Arbeit 1 wurde die Prädiktion von Studienerfolg nicht nur mit Persönlichkeitsmaßen auf Facettenebene sowohl von Fremd- als auch Selbst-Ratings untersucht, sondern auch der Einfluss von faking auf die Kriteriumsvalidität der Persönlicheitsfacetten. Ergebnisse konnten zeigen, dass Fremd-Ratings über Selbst-Ratings und Intelligenz hinaus Studienerfolg inkrementell prädizieren. Darüber hinaus konnte gezeigt werden, dass Faking die Kriteriumsvaliditäten auf Facettenebene in unterschiedlicher Weise beeinflusst, was einen sorglosen Umgang mit Faking verbietet. Arbeit 2 untersuchte den Einfluss der unterschiedlichen Abstraktionsebene von Prädiktor und Kriterium auf die Kriteriumsvalidität im Feld von Leistungsmotivation in der Schule. Dazu wurden Skalen zu Leistungsmotivation sowohl in einer Mathematik-spezifischen Formulierung als auch in einer globalen Formulierung Schülern zur Beantwortung vorgegeben. Diese Skalen dienten dann als Prädiktoren für Noten in Mathe, Physik und Deutsch. Durch Verwendung eines Multi-Trait-Multi-Method Ansatzes konnte die Varianz in diesen Skalen zerlegt werden. Ergebnisse zeigten, dass die Mathe-spezifischen Skalen durchgehend ein Plus an Varianz enthalten, welches unabhängig ist von der Varianz, die auf die einzelnen Motivationskonstrukte zurückgeht. Dies lässt den Schluss zu, dass domänen-spezifische Skalen entweder ein engeres Kontrukt von Leistungsmotivation messen (hier: mathe-spezifische Leistungs-motivation) oder, wahrscheinlicher, ein zusätzliches Konstrukt mitmessen. Dies wird untermauert durch den durchgängigen, positiven Zuwachs an Varianz unabhängig von der positiven oder negativen Valenz der Skalen. Das Korrelationsmuster zwischen der domänen-spezifischen Varianz und den drei untersuchten Noten legt außerdem den Schluss nahe, dass es sich bei diesem zusätzlichen Konstrukt um Selbstkonzept handelt. Arbeit 3 baute auf den bisherigen Ergebnissen auf und untersuchte die Konstrukt-validität von den Big 5 und möglichen sog. higher-order factors, nach Kontrolle von möglichen Verzerrungen (biases). Dazu mussten Versuchspersonen Selbst- und je zwei Fremdeinschätzungen von sich auf den Big 5 liefern. Durch Verwendung des jüngst entwickelten CTCM-1 Ansatzes konnten die Big Five ohne Rater-spezifischen Bias modelliert werden. Ergebnisse zeigten, dass die bias-bereinigten Big 5 Maße die Annahme eines higher-order factors wenig plausibel machen. Darüber hinaus konnte ein solcher potentieller Faktor nicht theoriekonform die positive Eigenschaft Intelligenz prädizieren.

Insgesamt verdeutlicht dies erneut die Problematik des Einflusses von unterschiedlichen Quellen und Verzerrungen auf die Kriteriumsvalidität von häufig eingesetzten Persönlichkeitsmaßen.

Abstract

This dissertation about the prediction of performance is based on three articles. Article 1 analyzes the prediction of academic performance by use of self-ratings, other-ratings and faked-ratings of personality measures not only on domain level but also on facet level. These three different scores were used to compare their influence on criterion validity. Result showed that other-ratings yield incremental validity above and beyond self-ratings and intelligence. Moreover, against prior findings for domain-level, faking does influence criterion validity on facet-level, with the influence not being uniform in direction. This result prohibits light-headed handling of faking. Article 2 analyzed the influence of different levels of abstraction of predictor and criterion in the realm of achievement motivation in school. For that, various achievement motivation scales were administered both in a global and a math-specific wording. These scales later on served as predictor for grades in math, physics and German. By modeling this data in a Multi-Trait-Multi-Method structural equation model different sources of variance could be disentangled. Results showed that math-specific scales are the better predictors. More so, these domain-specific scales have uniformly an increase in variance regardless of the positive or negative valence of the various achievement motivation scales. This leads to the conclusion that math-domain-specific scales either measure a narrower construct or, more probable, they tap an additional construct. This is backed by the uniform positive additional variance. Moreover, test-criterion correlation-pattern between the math-domain-specific variance and the three different grades makes it plausible that the additional construct tapped in these scales is self-concept. Article 3 built on these results and analyzed the construct-validity of personality's Big 5 and their possible higher order factor after controlling for singular rater biases. For that, self- and other ratings were obtained from participant. By use of the recently developed CTCM-1 approach, it was possible to model the Big 5 singular-rater-bias free. Results showed that these bias free Big 5 make the assumption of one higher order factor implausible. Moreover, such a factor would not uniformly predict intelligence as is claimed by advocates of this factor.

All in all, results emphasize the problem of influence of different sources and biases on criterion validity of well-established measures of personality.

1. Introduction

Prediction of behavior has long been and still is one of the pillars of psychology (Watson, 1913). Even more so, in the classic psychological triad of understanding/explaining, predicting and intervening, the aspect of predicting is the ultimate test for theories if they are working or not, for constructs, if they exist or not, for plans of action, if they are necessary or not. Hence, researchers work hard to bring by evidence that their predictions are valid. However, sometimes they fail, and predictions turn out to be weak. If something like this happens, the search for the underlying causes of this failure should begin. In other cases, predictions are not robust, i.e. they are good in some cases and not in others. Here, searching for the underlying causes is also necessary. In any search it proves to be useful to know exactly what is searched for. To acquire such knowledge, re-searchers should take a close look at what constitutes the integral parts of any prediction: that is, the predictor and the criterion. And even more important: there should be an answer of what are the defining parts of a predictor and a criterion and their interplay.

Unfortunately, researchers often settle with the conclusion that their predictors are fine if they just “work”. In other words, if predictions are not so bad, i.e. the criterion validities are not so low, they donot care where the predictive validity comes from. From a pragmatic point of view, this may be fine. But one part of the psychological triad is the understanding and explaining of phenomena. Hence, a good prediction without understanding where it stems from should not be an excuse to refrain from further investigations because it gives no insight into the mechanism that drives the prediction in order for it to work. In the worst case, this can lead to wrong conclusions and also wrong interventions, which could possibly do more harm than good. It should therefore be regarded as highly important to thoroughly investigate the inner workings of psychological predictions.The present set of studies aimed specifically at that: looking at different influencing aspects of predictions in order to better understand how predictions work.

One of the most important areas predictions are needed and applied in, is the area of performance. Performance is the indicator of how someone is doing with regards to certain standards or goals or reference groups. Performance can be assessed in specific institutions, for example school (e.g., Bratko, Chamorro-Premuzic, & Saks, 2006) or academic institutions (e.g., Poropat, 2009), at the job (Schmidt-Atzert, Deter, & Jaeckel, 2004), but also in other prominent areas like sports (e.g., Sulloway & Zweigenhaft, 2010). Psychology, of course, limits its research to those areas where at least some of the predictors are psychological constructs. One of the most important tasks in such research is finding those predictors who give best results while being, hopefully, reliably and validly assessable. But, as pointed out before, the pinpointing of certain predictors should not be the final goal, instead, understanding what drives the predictions, and looking for possible influences must not be forgotten.

2. The prediction of performance

The use of intelligence tests and personality questionnaires as predictors of performance has been practiced nearly since the first appearance of these concepts and tests (Webb, 1915). The criterion validity of these predictors has been proven by numerous studies. However, most of these studies stopped after reporting the size of the criterion correlation, and many lacked a

systematic approach to factors influencing and moderating the prediction itself. Often, they simply looked for the best predictor out of a set of possible predictors.

For example, Barrick and Mount (1991) could show in their meta-analytical approach that at least some measures of personality, namely some domains of the Big Five (Goldberg, 1990), yielded valid predictions of performance. Their analysis was limited to job performance, hence the three different criteria used were job proficiency, training proficiency, and personnel data. Additionally, Barrick and Mount only analyzed the Big Five domains without descending onto facet level. Furthermore, they did not touch onto the problem of potential moderators of criterion validity like the influence of social desirability responding (Murphy & Davidshofer, 2001; Paulhus, 2002). Nevertheless, their meta-analysis was strong evidence that personality measures make for good predictors of job performance. Despite these convincing results and the huge impact the meta-analyses has had on the scientific community (Mount & Barrick, 1998), the authors do not provide elaborated theoretical explanations for the mechanism causing the predictions.

In the realm of academic performance, a meta-analysis by Poropat(2009) aggregated studies showing the predictive validity of personality measures, especially of the Big Five. In particular, these studies showed the incremental predictive validity of personality measures above and beyond intelligence. Most of these studies were limited to the domain level. However, in the wake of Paunonen and Ashton (2001) who had shown that facets had higher criterion validity than domains with regard to over 40 criteria, Lounsbury and colleagues (Lounsbury, Sundstrom, Loveland, & Gibson, 2002) could show that some narrower personality facets (namely aggression, optimism, tough-mindedness, and work drive) are more powerful predictors of academic performance than the broader domains. Unfortunately, they did not include more facets into their study. And, as before, the mystery of the inner workings of the predictions, possible moderators and influencing factors were left untouched in most of these examples. All in all, despite the criticism all these studies show that predicting performance in diverse fields is possible.

2.1 The levels of abstraction that influence a prediction

However, the works by Paunonen and Ashton (2001) and Lounsbury and colleagues (2002) did bring to attention the problem of different levels of generalization with regard to predictors and criteria. Already Brunswick (1955) had pointed out that, for a good prediction, symmetry level has to be heeded. This means that predictor and criterion need to be on the same level of generalization. By this, it can either be understood that predictor and criterion have to be on the same level of abstraction or that predictor and criterion should be part of the same underlying domain.This is traditionally referred to as the rationale behind Brunswick's lens model. Both studies, the one by Paunonen and Ashton (2001) and the one by Lounsbury and colleagues (2002), however, did not systematically examine the difference in predictive validity by varying degree of specificity. That is, they used facets of the Big Five for their predictions in comparison to the Big Five domains, but they did not look where the better prediction came from.

The problem of specificity was even better acknowledged for the academic school setting: it was thought that different topics not only could promote topic-specific performance but also be the consequence of topic-specific predictors and thus driving forces of such performance. This is especially true for research in the realm of achievement motivation. Achievement motivation has been known to be a good predictor of academic performance (e.g., Nicholls,

1984). In recent years, the importance to differentiate between different domains has helped to find that domain specific achievement motivation yields better predictions of domain-specific, i.e. topic-specific, performance than global achievement motivation (Steinmayr & Spinath, 2007). This so-called domain-specific approach, which proved to be fruitful in school context not only for motivations, but also emotions (Goetz, Frenzel, Pekrun, & Hall, 2006; Goetz, Frenzel, Pekrun, Hall, & Lüdtke, 2007) and academic self-concept (Marsh, 1992, 1993), could show that one important influence on the quality of predictions apparently was level of symmetry. However, it still remained unclear whether a domain specific motivation was indeed domain specific with regard to motivation. Put differently, did domain specific *motivation* drive the prediction of performance or something different from motivation that becomes salient through the way the questions are asked within the domain specific questionnaires?

2.2 The sources of information that are used to build a prediction

When considering the issues discussed so far, which without a doubt only represent a small sample, it becomes apparent that there are many possible influences on any prediction. So far, the issues described could be regarded as being related to the questionnaires themselves, i.e., they were related to level of symmetry and to construct validity. The latter, however, is also related to the participants or better: to those who give the answers on a questionnaire. In the way those answers are given, the validity of the construct is formed. Construct validity, of course, will directly influence the prediction. But such a statement is trivial. Not so trivial is the fact that participants are not always able or willing to give the most appropriate answer. When participants are asked to give ratings about themselves or others, there always looms the possibility that their assessments are skewed. They could, for example, alter their ratings (i.e., fake) in order to deceive others or to deceive themselves. Even if they give the most accurate and truthful assessment possible to them such an assessment might suffer from their obstructed perspectives. They just might not know better because they did not have access to vital information to give a more accurate assessment. Hence the question: what is such an answer worth? What conclusions can be drawn from such an answer? Which predictions made? Because faking is known to influence construct validity of personality measures (Pauls & Crost, 2004), it is sensible to examine whether it also influences criterion validity related correlations. For the domain level, this has been done before meta-analytically and experimentally with apparently positive, i.e. encouraging results (Ones, Viswesvaran, & Reiss, 1996; Ziegler & Bühner, 2009). However, the level of specificity had not been regarded in those studies. Results only applied to the domain level, not facet level. Therefore, a conclusive answer regarding criterion validity is still missing.

Regardless of how faking can influence self-ratings and therefore distort answers, the quest for minimizing the influence of faking should be regarded as highly important. Distorting ones' answers from the "true" score is commonly referred to as bias. With stand-alone, i.e. single self-ratings, biases are hard to detect and control for (C. DeYoung, 2010). One of the approaches advocated, therefore, is the use of multi-rater data (Anusic, Schimmack, Pinkus, & Lockwood, 2009; Biesanz & West, 2004; C. G. DeYoung, 2006). Such an approach is inherently intertwined with the use of other-ratings, which, for personality data, most of the time are peer-ratings. Whereas the use of multi-rater data in order to minimize biases is not new (Biesanz & West, 2004; C. G. DeYoung, 2006), only recently developed methodological approaches allow the correct modeling of such data. The novelty of these approaches lies in the way data from different raters is treated: other-ratings are not independent, but they are nested into one specific target, i.e. the object the rating is given for. Studies up until now did

not take this nestedness of multi-informant-data into account. Therefore, conclusions drawn from these studies are weak, to say the least, and the question what happens with construct validity when biases are controlled for still remains.

All this taken together, the following questions ensue regarding the prediction of performance using measures of personality constructs:

1. Do different sources of information, i.e. classes of raters, influence, boost and round off the prediction of performance?
2. Does information given with the intention to fake influence construct validity and through that the prediction, i.e. the criterion correlation, on facet level?
3. Does the heeding of symmetry level between predictor and criterion yield better predictions, as has been shown before? If so, what is driving the better prediction on a domain-specific level and what are the inner workings of such a predictor with regard to construct validity?
4. Does construct validity change for personality measures when multi-informant data is modeled in such a way as to minimize the influence of biases and to control for the nestedness of data?

The questions outlined above where the starting point for the three articles that are the base of the current dissertational project. Research for all three articles has been conducted while working at the chair of Psychological Assessment, held by Prof. Dr. Matthias Ziegler, at the Psychological Institute of the Humboldt Universität zu Berlin. Each article tried to tackle a different problem regarding the prediction of performance.

Article 1 looked for the influence of different classes of raters on the prediction and at the same time at the problem of faking for criterion validity on facet level. For that, studies in article 1 used the Big Five to predict academic performance, i.e. performance in an exam, after controlling for intelligence. (questions 1 & 2)

Article 2 looked at the problem of symmetry level between predictor and criterion when predicting domain-specific school grades with global and domain-specific measures of achievement motivation. In this study, variance decomposition was used to better understand the inner workings of the predictor and by that to take a closer look at how domain-specific wording affects achievement motivation scales. (question 3)

Article 3 then further investigated the issue of construct validity with regard to the influence of biases. By controlling for biases it should be analyzed whether the emergence of specific constructs is due to biased data and even more so whether specific test-criterion correlations are substantive or also just effects of biases. In this specific case, it was investigated whether a general factor of personality (GFP) as recently propagated is more than a chimera and whether such a factor can predict performance on an intelligence test.

In the following section, the results of each article are presented in a short overview.

3. Summary and results of the three articles:

3.1 Study 1: Predicting Academic Success with the Big 5 Rated from Different Points of View: Self-Rated, Other Rated and Faked

Study 1 investigated data from 145 undergraduate students who had to give personality ratings for the Big Five once under neutral instructions and once under the instruction to apply for a psychology university program. In the latter instruction it was made clear to the participants that being accepted or not depended solely on how they answered the questionnaire. This instruction would promote faking in order to being accepted by the program. In addition to the two self-ratings, each participant had to provide ratings on her/himself given by two peers. All of these measures served as predictors. As a criterion, we used grades in a statistics exam two month after the personality ratings were obtained. Because intelligence has been shown to be one of the best predictors of performance intelligence scores were also obtained in order to replicate these findings and to control for the effect of intelligence when using personality measures as predictors.

Results confirmed the role of intelligence in the prediction of academic performance. Furthermore, results replicated prior findings that other-ratings yield incremental validity to self-ratings. Because the personality measures used included facet scores it could be shown that it is selected facets driving the test-criterion correlations. The descent on facet level also gave new insight into the effect of faking: faking, as had to be expected per definition (Ziegler & Bühner, 2009), did not occur uniformly for all facets and not even in the same direction for all facets it occurred. Because of that, criterion validity suffered on facet level whereas on domain level prior findings of unaltered criterion validity could be replicated. This, however, could only be possible because correlation coefficients for some facets increased while they decreased for other facets.

All in all, this study could show that both self- and other ratings make unique contributions, not shared by each other, to the prediction of academic performance. Furthermore, results stressed the importance to look not only at domain level but also at the underlying facet level when investigating criterion validity and before making claims about the influence of biases, in general, regarding predictions of performance. Moreover, the inner workings of criterion validities estimated in different situations were elucidated.

These results made it quite clear that different levels of abstraction and hence different degrees of symmetry influence predictions. This has been known for quite a while in the scholastic context where domain-specific measures are employed regularly. These domain-specific measures also regularly yield better predictions. But the mechanism behind the improvement in predictions has been, so far, left alone. This fact was taken as a starting point for study two: it set out to investigate the mechanism of domain-specific predictors in a realm where those predictors were most established, the scholastic context.

3.2 Study 2: Global Versus Specific Approaches to Studying Achievement Motivation: An MTMM Study

In Study 2, three hundred twenty-five school children gave ratings on different measures of achievement motivation, which had been changed in their wording in order to once reflect a global, unspecific motivation, and once a domain-specific motivation. Because math is generally regarded as an important and sometimes emotionally loaded school topic, it was chosen as the domain-specific topic. These measures, then, were used as predictors of school grades in three different subjects: math, physics, and German. By doing so, it was possible to compare predictions by the global measures with predictions by the math-domain-specific measures for the same underlying trait. Furthermore, because three criteria were available, comparisons could be made between a matching domain criterion, i.e. math, and non-matching domain criteria, i.e. physics and German. Furthermore, to elucidate the workings of the predictors, we used a multi-trait multi method (MTMM) approach by structural equation modeling (SEM) to decompose the different sources of variance that should make up each measure. Apart from the variance due to specific approaches to achievement motivation (i.e., mastery, performance, approach, avoidance, hope for success, fear of failure), it should also be possible to account for the variance due to different wording of the various measures (i.e. the global, unspecific wording and the math domain-specific wording).

Results confirmed the prior found superiority of domain-specific measures as long as the criterion matched the domain. Accordingly, the math-domain-specific measures yielded better predictions of grades in math, but not so of physics or German grades. For non-matching criteria, the predictions of math domain-specific measures of motivation were no different from global measures of motivation.

Through the MTMM approach, the different sources of variance could be decomposed. It could be shown that the variance due to motivational constructs did not differ between the global and the domain-specific measures. This means that there was an equal amount of variance in both classes of measures due to mastery, performance, approach, avoidance, hope for success, and fear of failure. However, communalities for the domain-specific measures were higher. This surplus could be located in the variance due to the math-domain-specific wording. Interestingly, after adding grades to this structural equation model, it was mostly this variance due to domain-specific wording driving the better prediction of grades in the domain-matching subject, in comparison to the variance also found in the global measures. All these findings held true even after controlling for conscientiousness.

These results gave rise to two possible explanations. First, it could be assumed that these domain-specific measures capture a narrower facet of achievement motivation. In this case: math specific achievement motivation. Such a motivation, however, would be not so much qualitatively different from global achievement motivation but just an add-on or a hierarchically lower level trait. This seems implausible considering that at least fear of failure and avoidance have a negative valence with respect to the other constructs. The additional variance, however, was positively found in all scales. Furthermore, the latent factors of mastery, performance, approach, avoidance, hope for success, and fear of failure did capture the same amount of variance regardless of the wording of the measures pointing at the fact that the core of the motivational constructs stayed the same. All taken together, makes this explanation, while not completely groundless, not as plausible as the second explanation. That explanation argues that the additional variance in the domain-specific measures could be

attributable to a second, additional construct. Such an explanation can more easily be aligned with the fact that regardless of the valence of the underlying scale, the domain specific wording added variance. Furthermore, all other components retained the same amount of variance. Plausibly, this additional construct could be self-concept. Such an explanation is warranted by the pattern of the test-criterion correlations of the additional variance: the variance due to the math domain-specific wording yielded a positive correlation with math grades, a smaller positive correlation with physics grades and a negative correlation with German grades. The negative correlations reflect findings from self-concept research: A positive math self-concept has a negative influence on language grades and vice versa (Marsh, 1986, 1990; Möller & Köller, 2004; Schilling, Sparfeldt, Rost, & Nickels, 2005). Because no self-concept questionnaire was included in the study, the soundness of this explanation could not be confirmed conclusively. This remains a task for further studies.

Reminded by these findings that sometimes only very specific variance will drive a test-criterion correlation, I turned to a recently very controversial topic, namely the topic of a general factor of personality (GFP), which should drive people on its positive pole to success and greater fitness in comparison to people on the other end of its dimension. Because advocates of the GFP base its existence on differential *K* theory(Rushton, 1985), they argue that the GFP predicts generally positively valenced traits like intelligence or agreeableness and conscientiousness, higher emotional stability and so forth.

3.3 Study 3:Really a Single Factor of personality? A Multi-Rater approach to the GFP and below

Taking into account results from Study 1, namely that self- and other ratings do not completely overlap, but are valuable sources of information when it comes to personality, and from Study 2, namely that for certain test-criterion correlations only a small amount of the whole variance is drivingsuch correlations, Study 3 set out to examine whether the variance in the GFP is due to bias and if so whether predicted test-criterion correlations between the GFP and traits like intelligence would still to be found after controlling for the influence of biases. Of course, this more or less is also a direct investigation of the influence of source and bias on the construct validity of Big 5 questionnaires.

As has been mentioned above, up until now, multi-rater approaches suffered from the fact that the nestedness of the data was not taken into account. The newly developed CTCM-1 approach by Eid and colleagues (Eid et al., 2008) allows to do just that: take nestedness into account and control for different rater biases.

N=404 undergraduate students were recruited who in addition to their self-ratings on a Five Factor Model questionnaire (Borkenau & Ostendorf, 1993) had to provide ratings on themselves by two peers. Additionally, for use later on as a criterion the Intelligence Structure Test 2000-R (Amthauer, Brocke, Liepmann, & Beauducel, 2001) which provides scores for verbal, numerical and figural intelligence as well as for reasoning was administered. With these data composed of self- and other-ratings not only the Five Factors of personality according to the CTCM-1 approach were modeled but also above these five factors the GFP. Because of using the CTCM-1 approach, all five personality factors were free of individual rater biases. The only substance ending up in these factors was the shared variance by all three raters. Apart from variance due to the construct being rated, for example extraversion, variance due to bias could only be existent if all three raters exhibited the same bias. In any case, with these bias adjusted data the GFP did not exist. Variance of a possible GFP did not

reach statistical significance. Of course, this might have been due to power issues, which is unlikely given the sample size but cannot be ruled out completely.

In a second step we wanted to test the prediction made by Differential *K* theory, namely that the GFP is a predictor of positively valenced traits. In order to do so, we therefore added measures of intelligence to our model and used it as criterion. Interestingly, the variance inside the GFP correlated positively with verbal intelligence but negatively with numerical and figural intelligence. Apparently, whatever is captured inside the GFP does not positively predict all possible positive traits. Numerical and figural intelligence are from an evolutionary perspective by no means less important than verbal intelligence. Even more so, verbal intelligence most certainly is much later evolved than figural intelligence. A general positive influence of the GFP could therefore be ruled out. But what could the variance inside the GFP then be? As a possible hypothesis impression management was forwarded. Apparently, the GFP had positive loadings on traits agreeableness, conscientiousness and emotional stability. These are traits welcomed by society. In addition to that, verbal intelligence was the only positively correlated facet of intelligence with the GFP. These findings taken together, the hypothesis was forwarded that whatever variance is inside the GFP, it is generated by a positive overlap of the views different sources have on one person's personality. Such an overlap is most easily achieved by interactional behavior but also by "story telling", i.e. by verbally conveying one's own positive traits.

4. Conclusion

Above I asked four different questions which I set out to answer through the 3 articles presented as a dissertational project here. The first question asked whether different sources of information, i.e. classes of raters, can yield better predictions of performance. This has been positively answered by the study in article 1. Other-ratings provide an increment in predictive validity to self-ratings and intelligence. Even though this had been shown in prior studies, this was the first study to investigate on facet level in an academic setting while controlling for intelligence at the same time.

The second question dealt with the problem whether the intention to fake influences criterion validity on facet level. It could be shown in article 1 that on facet level, faking does influence the criterion validity. This influence sometimes leads to an increase, sometimes to a decrease of test-criterion correlations. This problematic finding was in contrast to the encouraging findings for criterion-validity on domain level and could be a promising start for future research on the impact of faking.

With the third question I wanted to investigate what the underlying mechanisms are for the better predictions on a domain-specific level with regard to domain-specific criteria. It could be shown that in domain-specifically worded measures of achievement motivation an additional source of variance could be found in comparison to the same measures of achievement motivation when phrased in a global manner. These results gave rise to the hypothesis that this additional variance was not so much due to the measuring of a narrower construct, but due to an additional source of variance. As a possible candidate for such a source, self-concept could be identified. This hypothesis was backed by the findings that this additional variance correlated positively with the math grade (i.e., the matching domain-specific school topic) but negatively with the German grade (i.e., a non-matching domain-specific school topic). Because this correlation pattern mirrors the patterns found in self-concept research, such an explanation is thought to be highly plausible. However, only future studies can provide conclusive answers.

Question 4, finally, asked whether construct validity for personality measures changes when multi-informant data is modeled in such a way as to minimize the influence of biases and to control for the nestedness of data. By using the newly developed CTCM-1 approach I could show that a construct like the General Factor of Personality (GFP) is most likely driven by variance due to bias. The prediction this GFP should be able to make, i.e. predictions of intelligence and positively valenced personality traits like agreeableness, could also not be found uniformly. Instead, through an incongruent correlation pattern of the GFP with facets of intelligence, the hypothesis was forwarded that the GFP is mostly due to successful impression management.

Summing up, the three studies provide insight into the impact different sources of information, biases, and levels of abstraction have on the criterion validity of widely used personality questionnaires. Because these different aspects apparently can sometimes drive test-criterion correlations, it is vital to control for them as much as we can before we take criterion validity as adequate and predictions as valid as it is sometimes done. Only a sensible and careful approach to data, which ultimately will be the base for our prediction, will give correct and truthful and sound results.

References:

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). I-S-T 2000 R (Intelligenz-Struktur-Test 2000 R) [Intelligence-Structure-Test 2000 R]. Göttingen: Hogrefe.

Anusic, I., Schimmack, U., Pinkus, R. T., & Lockwood, P. (2009). The Nature and Structure of Correlations Among Big Five Ratings: The Halo-Alpha-Beta Model. Journal of Personality and Social Psychology*, 97*, 1142-1156.

Barrick, M. R., & Mount, M. K. (1991). The BIG Five Personality Dimensions and Job Performance: A Meta-Analysis. Personnel Psychology*, 44*, 1-26.

Biesanz, J. C., & West, S. G. (2004). Towards Understanding Assessments of the Big Five: Multitrait-Multimethod Analyses of Convergent and Discriminant Validity Across Measurement Occasion and Type of Observer. Journal of Personality*, 72*, 845-876.

Borkenau, P., & Ostendorf, F. (1993). NEO-PI-R nach Costa und McCrae [NEO-PI-R by Costa and McCrae]. Göttingen: Hogrefe.

Bratko, D., Chamorro-Premuzic, T., & Saks, Z. (2006). Personality and school performance: Incremental validity of self- and peer-ratings over intelligence. Personality and Individual Differences*, 41*, 131-142.

Brunswik, E. (1955). Representative Design and Probabilistic Theory in a Functional Psychology. Psychological Review*, 62*, 193-217.

DeYoung, C. (2010). Toward a Theory of the Big Five. Psychological Inquiry*, 21*, 26-33.

DeYoung, C. (2006). Higher-order factors of the Big Five in a multi-informant sample. Journal of Personality and Social Psychology*, 91*, 1138-1151.

Eid, M., Nussbeck, F., Geiser, C., Cole, D., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. Psychological Methods*, 13*, 230-253.

Goetz, T., Frenzel, A. C., Pekrun, R., & Hall, N. C. (2006). The domain specificity of academic emotional experiences. Journal of Experimental Education*, 75*, 5-29.

Goetz, T., Frenzel, A. C., Pekrun, R., Hall, N. C., & Lüdtke, O. (2007). Between- and within-domain relations of students' academic emotions. Journal of Educational Psychology*, 99*, 715-733.

Goldberg, L. (1990). An alternative “description of personality”: The Big-Five factor structure. Journal of Personality and Social Psychology*, 59*, 1216-1229.

Lounsbury, J. W., Sundstrom, E., Loveland, J. L., & Gibson, L. W. (2002). Broad versus narrow personality traits in predicting academic performance of adolescents. Learning and Individual Differences*, 14*, 65-75.

Marsh, H. W. (1986). Verbal and Math Self-Concepts – an Internal External Frame of Reference Model. American Educational Research Journal, *23*, 129-149.

Marsh, H. W. (1990). The Structure of Academic Self-Concept - the Marsh Shavelson Model. Journal of Educational Psychology, *82*, 623-636.

Marsh, H. W. (1992). Content Specificity of Relations between Academic-Achievement and Academic Self-Concept. Journal of Educational Psychology, *84*, 35-42.

Marsh, H. W. (1993). The Multidimensional Structure of Academic Self-Concept - Invariance over Gender and Age. American Educational Research Journal, *30*, 841-860.

Möller, J., & Köller, O. (2004). Die Genese akademischer Selbstkonzepte: Effekte dimensionaler und sozialer Vergleiche [On the development of academic self-concepts: The impact of social and dimensional comparisons]. Psychologische Rundschau, *55*, 19-27.

Mount, M. K., & Barrick, M. R. (1998). Five reasons why the "big five" article has been frequently cited. Personnel Psychology, *51*, 849-857.

Murphy, K. R., & Davidshofer, C. O. (2001). Psychological Testing (Vol. 5th): New Jersey: Prentice Hall.

Nicholls, J. G. (1984). Achievement-Motivation - Conceptions of Ability, Subjective Experience, Task Choice, and Performance. Psychological Review, *91*, 328-346.

Ones, D. S., Viswesvaran, C., & Reiss, A. D. (1996). Role of Social Desirability in Personality Testing for Personnel Selection: The Red Herring. Journal of Applied Psychology, *81*, 660-679.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson & D. E. Wiley (Eds.), The role of constructs in psychological and educational measurement (pp. 49-69): Mahwah, NJ, US: Lawrence Erlbaum Associates.

Pauls, C. A., & Crost, N. W. (2004). Effects of faking on self-deception and impression management scales. Personality and Individual Differences, *37*, 1137-1151.

Paunonen, S. V., & Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. Journal of Personality and Social Psychology, *81*, 524-539.

Poropat, A. E. (2009). A Meta-Analysis of the Five-Factor Model of Personality and Academic Performance. Psychological Bulletin, *135*, 322-338.

Rushton, J. (1985). Differential K Theory: The sociobiology of individual and group differences. Personality and Individual Differences, *6*, 441-452.

Schilling, S. R., Sparfeldt, J. R., Rost, D. H., & Nickels, G. (2005). Facets of academic self-concept - Validity of the Differential Self-Concept Grid (DISC-Grid). Diagnostica, *51*, 21-28.

Schmidt-Atzert, L., Deter, B., & Jaeckel, S. (2004). Prädiktion von Ausbildungserfolg: Allgemeine Intelligenz (g) oder spezifische kognitive Fähigkeiten? Zeitschrift fur Personalpsychologie*, 3*, 147-158.

Steinmayr, R., & Spinath, B. (2007). Predicting school achievement from motivation and personality. Zeitschrift für Pädagogische Psychologie*, 21*, 207-216.

Sulloway, F. J., & Zweigenhaft, R. L. (2010). Birth Order and Risk Taking in Athletics: A Meta-Analysis and Study of Major League Baseball. Personality and Social Psychology Review*, 14*, 402-416.

Watson, J. B. (1913). Psychology as the behaviorist views it. Psychological Review*, 20*, 158-177.

Webb, E. (1915). Character and intelligence: An attempt at an exact study of character. British Journal of Psychology Monographs *1*, 1-99.

Ziegler, M., & Bühner, M. (2009). Modeling Socially Desirable Responding and Its Effects. Educational and Psychological Measurement*, 69*, 548-565.

Predicting Academic Success with the Big 5 Rated from

Different Points of View: Self-Rated, Other Rated and Faked

Veröffentlicht in:

Ziegler, M., Danay, E., Schölmerich, F., & Bühner, M. (2010). Predicting academic success with the Big 5 rated from different points of view: Self-rated, Other rated and faked. *European Journal of Personality*. doi: 10.1002/per.753

Global Versus Specific Approaches to Studying Achievement Motivation: An MTMM Study

Abstract

Achievement motivation has been shown to be an important factor in predicting academic performance. Particularly when phrased in domain-specific (e.g., math) language, achievement motivation measures yield better predictions in this domain than globally phrased measures. In order to investigate what accounts for the difference of these domain-specific versus global scales, 325 school children assessed themselves once on globally and once on domain-specifically phrased scales of different operationalizations of achievement motivation.Grades in three subjects served as criteria. The differently phrased scales were compared in terms of reliabilities, means, and intercorrelations, and then subjected to an MTMM analysis using SEM. Results showed higher method factor loadings for the domain-specific scales. Test-criterion correlations for the domain-specific scales were driven mainly by this method variance. Two possible explanations for the source of this variance are discussed: Either the specific measures capture a narrower facet such as math motivation or else the self-concept is responsible.

Keywords: achievement motivation, anxiety, domain specificity, emotions, self-concept, MTMM, criterion validity

Global Versus Specific Approaches to Studying Achievement Motivation: An MTMM Study

There has been abundant research individuating motivational aspects (Covington, 2000; Dweck, 1986) and personal expectancy (Schunk, 1991; Weiner, 1985) as predictors of academic achievement above and beyond intelligence (Nicholls, 1984; Steinmayr & Spinath, 2007, 2009).

Whereas intelligence has undeniably been found to be the most powerful predictor of academic achievement (Kuncel, Hezlett, & Ones, 2001), it has also been shown to be quite resistant to intervention (Campbell & Ramey, 1994; Perkins & Grotzer, 1997). Therefore, from a pedagogical point of view, the primary focus of achievement research has to be put on other classroom factors that interventions can have an effect on; above all, motivation and emotion. However, initial programs to lever motivation have yielded only moderate results. The possibility that motivation is not consistent across situations was picked as one reason. Hence, researchers have taken into account that motivation and emotion should vary according to the situation (i.e., the specific content domains as sources of these specific emotions and motivations). Thus, the argument becomes that the prediction of academic achievement in a certain subject would best be done by using predictors that relate to this specific subject. This is backed by Brunswik's lens model approach (1955). In psychology, there have been numerous examples corroborating this assumption. In personality, lower-order facets tend to be better predictors than the according higher-order dimensions if the criterion is very specific (Bagby, Costa, Widiger, Ryder, & Marshall, 2005; Paunonen & Ashton, 2001; Ziegler, Danay, Schölmerich, & Bühner, 2010; Ziegler, Knogler, & Bühner, 2009).

In academic contexts, this so-called domain-specific approach has proven to be advantageous to a global approach in explaining grades (Steinmayr & Spinath, 2007). Moreover, there has been research on the domain specificity of achievement motivation (Bong, 2001; Green, Martin, & Marsh, 2007; Martin, 2008; Wigfield, 1997; Wigfield, Guthrie, Tonks, & Perencevich, 2004) and emotions (Goetz, Frenzel, Pekrun, & Hall, 2006; Goetz, Frenzel, Pekrun, Hall, & Lüdtke, 2007), particularly of anxiety (Meece, Wigfield, & Eccles, 1990) and the self-concept (Marsh, 1992, 1993).

Although research supports the notion that these domain-specific measures really are the better predictors of academic achievement compared to global measures, besides Brunswik's lens model idea, there still has been no clear conceptualization of what it is inside these measures that yields better predictions. The importance of understanding this mechanism has already been expressed by Finney and colleagues (Finney, Pieper, & Barron, 2004) who explicitly called for a direct comparison of course- versus domain-specific measures in order to understand the predictive validity of achievement-motivation measures. So far, only sparse research has been conducted on this. Steinmayr and Spinath (2009) have stressed the importance of differentiating between global and domain-specific measures when predicting scholastic achievement. However, although they included motivational measures in their study, they limited the domain-specific aspect of their study to ability self-concepts and values. Another study (Greene, Miller, Crowson, Duke, & Akey, 2004) did use domain-specific measures for all of their scales in predicting achievement in high school, but the only domain-specific measures they employed were of self-efficacy, cognitive engagement, and achievement. Thus, neither of the studies actually compared domain-specific and global measures of achievement motivation. Therefore, two questions remain: (a) are domain-

specific measures of motivation actually better predictors of achievement? and (b) if so, what is driving this advantage?

The current study set out to address these questions by disentangling the different variance components in global and domain-specific achievement motivation measures. By applying a multitrait-multimethod approach, the different variance sources (i.e., trait and method) were differentiated. This made it possible to find out where the better predictive validity originated from. Apart from that, in achievement motivation research, over time, different approaches have been developed, accentuating different aspects of the need for achievement. Based on these different conceptualizations, different ways of measuring motivation have evolved. Therefore, to cover a large area of what is thought to be part of achievement motivation, these different conceptualizations were included. Following is a short overview of these different approaches to the need of achievement motive.

Different Approaches to Achievement Motivation

Mastery and Performance

Achievement motivation was first introduced through a systematic approach into psychology by Murray with the coinage of need for achievement (1938). According to Murray, need for achievement constitutes an individual's drive to accomplish certain goals or meet standards of excellence. Regarding these goals, theorists developed a classification framework that distinguishes between two classes of goals: mastery and performance (e.g., Nicholls, 1984). Whereas mastery orientation drives a person to acquire knowledge and abilities simply for the sake of acquiring this knowledge or these abilities, a more performance-orientated person tries to outperform others and to do "better" than the rest, regardless of how good or bad his or her acquired abilities and understanding really are (Dweck & Leggett, 1988; Linnenbrink & Pintrich, 2002). Whereas this distinction has not always been as clear-cut (Bouffard, Boisvert, Vezeau, & Larouche, 1995; Bouffard et al., 1998), and these goals have not been as mutually exclusive (Elliot & Murayama, 2008) as one would tend to believe at first sight, it is well-established in the current literature on achievement motivation to conceive of mastery and performance goals as independent because they are rooted in different frames of comparison, namely, absolute versus normative.

Approach and Avoidance

Another important goal distinction inside the need for achievement framework what was found in the seminal work by McClelland and colleagues (McClelland, Atkinson, Clark, & Lowell, 1953) and emphasized by modern achievement goal theorists and integrated into their theory (Elliot & Harackiewicz, 1996; Harackiewicz, Barron, Pintrich, Elliot, & Thrash, 2002) is to be made between approach and avoidance strivings. One can put one's effort into approaching situations in which to employ these competences (approach) or into avoiding such situations (avoidance). The valence of these two strivings is antipodal and resides according to McClelland and colleagues in the affective experience in achievement situations. That is, whatever a person believes to be happening with her/himself during such situations determines the strength of approach and avoidance. If someone will be elated by a certain situation, he or she will try to seek it out; but if, on the other hand, someone will be humiliated by the same situation, he or she will try to avoid it.

Therefore, whereas mastery and performance can be seen as the "rational" parts of the need for achievement, approach and avoidance are heavily rooted in a person's affective experience. Elliot and McGregor (2001) integrated the dimensions of approach versus and avoidance and mastery versus performance into their achievement goal theory, and by conceptualizing an orthogonal relationship between them, adopted a 2x2 framework. Hence, there is mastery-approach and mastery-avoidance, performance-approach and performance-avoidance, which are assessed independently in the widely used Achievement Goal Questionnaire (AGQ) also used in this study.

There are some theoretical considerations regarding the constructs described so far: First, in their multiple-indicator-correlated trait-correlated method model, Elliot and Murayama (2008) allowed a correlation between latent approach and avoidance. Apparently, it is nearly impossible to employ an action of approach and avoidance at the same time. Therefore, there is some connection between approach and avoidance that we will also take into account in our model. Second, there seems to be an imbalance in predictive power between mastery-avoidance and mastery-approach goals. Mastery-avoidance "represent a puzzling motivational hybrid, and it simply is not clear how these two seemingly discordant components operate together in the process of goal regulation" (Elliot & Murayama, 2008, p. 625).It is conceivable that the mastery component in this goal is less salient than inside the mastery-approach goal since this definition of mastery - not to perform worse than before - takes its starting point from the "minimum" and seems to settle on a lower level. Accordingly, it has been found that mastery-avoidance goals tend to be misinterpreted as approach goals and, in general, happen to be employed rarely (Ciani & Sheldon, 2010). Therefore, the relationship between mastery and approach is possibly much stronger than that between mastery and avoidance (Finney, et al., 2004). To account for this, we again allowed a correlation between mastery and approach.

Fear of Failure and Hope for Success

In the framework of achievement motivation, the labels of fear of failure (FF) and hope for success (HS) have been in use since the establishment of the concept (Clark, Teevan, & Ricciuti, 1956). The notion behind these labels, however, is and was quite diverse. Whereas on the one hand, these concepts have been conceptualized as either needs or motives or affective tendencies(Conroy, 2003), these definitions are not concerned with the distinguishing mark to the notion of avoidance and approach. For example, Murray (1938) had used the term infavoidance to describe the avoidance of feelings of inferiority in comparison to one's peers. Therefore, he subsumed FF under the need to not feel inferior. Similarly, McClelland and colleagues (1953) linked FF and HS to the motives of approach (HS) and avoidance (FF). On the other hand, Clark and colleagues (1956)had already pointed out that the approach and avoidance motives are accompanied but not equal to hope for success and fear of failure. Accordingly, Heckhausen (1977, p. 309) denoted HS and FS as "two tendencies" within the achievement motive and therefore disentangled the incentive construct from the expectancy construct. There is an obvious theoretical similarity to approach and avoidance. McClelland and colleagues (1953) stressed the fact that approach and avoidance are linked to the emotions experienced during or after the achievement situation. Thus, they are not to be confounded with HS and FF because the latter two occur before the achievement situation. HS and FF are thus best described as hope or fear evoked by imagining a certain achievement situation and anticipating the expected emotional experience caused by success or failure. In Atkinson's words (1957): "The motive to avoid failure is considered a disposition to avoid failure and/or a capacity for experiencing shame

and humiliation as a consequence of failure" (p. 360). Because of the emotion evoked by the imagined achievement-situation outcome, HS/FF influence the definition of one's goal regarding this achievement situation (e.g., to avoid this situation or to approach it). Based on such a conceptual framework, Gjesme and Nygård devised a questionnaire tapping HS and FF as measures of the either positive or negative emotions the achievement situation should be loaded with (Gjesme, 1981; Gjesme & Nygård, 1970; Nygård & Gjesme, 1973). A version of this questionnaire was used in this study.

Measures of Personality in the Achievement Context

Achievement striving can also be seen as part of personality (Ziegler, Schmukle, Egloff, & Bühner, 2010). Recent meta-analyses by O'Connor and Paunonen (2007) and by Poropat (2009) have shown conscientiousness to be the one domain of personality to be most highly associated with academic performance after controlling for intelligence. This may partly be due to the facet of conscientiousness known as achievement striving (Ziegler, et al., 2009). Studies concerned with the facet structure of conscientiousness have repeatedly found a motivational component (MacCann, Duckworth, & Roberts, 2009; Roberts, Chernyshenko, Stark, & Goldberg, 2005). But even besides achievement striving, at least three other facets have been linked to performance (i.e., order, dutifulness, and self-discipline; Ziegler, et al., 2009). It was therefore sensible to include a measure of conscientiousness to control for its influence when analyzing any relationships between motivation and performance.

Global versus Domain-Specific Measures

When looking at the terms "global" and "domain-specific" in psychology, it seems that these have been used mostly as concepts denoting different levels of abstraction when referring to one particular topic. For example, someone can be punctual all the time (global) or just punctual when going to school (domain-specific). At a lower level of abstraction, someone can always be punctual when going to school (global) or only when going to math class (domain-specific). Hence, when defining something as "domain-specific," there is always some aspect limiting the generalizabilty. In general, the difference between global and domain-specific measures is analyzed usually from the consistency perspective of personality (Fleeson & Noftle, 2008). Nevertheless, it has been noted that adding up the different domain-specific measures does not equal the global measure (Rosenberg, Schooler, Schoenbach, & Rosenberg, 1995). The reasons behind this, however, have not been sufficiently investigated empirically until now. Nevertheless, with regard to Brunswik's lens model, it has been seen as vital to match the specificity[1] of the predictor to the specificity of the criterion. The underlying rationale here is that someone should "use a rifle to hit the center of a target" but "use a cannon to blast a large area" (Ironson, Brannick, Smith, Gibson, & Paul, 1989, p.200). From a theoretical point of view, this is quite clear and unambiguous. However, when defining scales and generating items, a researcher has to have in mind not only a single circumscribed area or construct to be measured, but also the means to set the correct distance between the different levels of abstraction. This distinction is most often made only on a theoretical level. The actual operationalization of this measure is usually relegated to a footnote (e.g., Rosenberg, et al., 1995), and the inner workings of the domain-specific measures are left alone.

[1]It should be noted here that, throughout this article, the term *specificity* is used in its strict basic sense and as the opposite of *global*. It is unrelated to the one in the terminus technicus pair *specificity-sensitivity* used to describe the psychometric properties of a test that is used to differentiate between certain groups.

Aim and Methodological Approach of the Study

Whereas there is no denying that specific measures are better predictors of specific achievement in an academic setting, the underpinnings at work have, until now, always been taken at face value or not discussed at all. Therefore, we set out to take a closer look and compared globally phrased scales of motivation with the same scales when phrased specifically for one school subject. For our study, we opted for math as the specific subject. Math has been known to have a special, and quite often negative, valence for many students because it is the subject they struggle with the most (Aiken, 1976; Ma, 1999).

In particular, we tried to disentangle different sources of variance inside the measures used for being able to see what ultimately has the best predictive power for achievement in a particular school subject. To this end, we first examined whether there were any differences in the reliabilities of these measures, in their intercorrelations, or in their means, in order to establish whether the constructs measured globally versus specifically are comparable or not, and whether subjects responded in a similar manner to these measures. In addition, we explored how the scales phrased in a global manner versus a domain-specific manner for mathematics were related to school performance (grades) in mathematics, but we also assessed how the scales were related to performance in physics (a domain adjacent to mathematics), and in German (a more disparate domain).

In a second step, we sought to analyze which sources of variance constituted the measures used in this study. For this, we used a latent multitrait-multimethod (MTMM) approach with several latent variables tracing the different approaches to the need for achievement described above (the 2x2 achievement goal framework and the fear of failure/hope for success differentiation) and also two method variables, namely, one for scales phrased globally and one for scales phrased domain-specifically. To compare the domain-specifically versus globally phrased scales, we looked at the loadings in the structural equation model and additionally compared the communalities (h^2) of each of the scales. This way it was possible to determine the amount of systematic variance that the traits and methods explained in each measure (Aiken & Groth-Marnat, 2006, p. 453). Finally, we were concerned about criterion-validity-related evidence to determine which sources of variance contributed to predicting grades and how much so. The correlations between school performance and the latent variables that depict different approaches to need for achievement on the one hand, and different levels of abstraction, on the other, were hypothesized to give a clearer picture of the inner workings in terms of both the power of the different theoretical approaches integrated in our study and the mechanisms underlying methodological factors of domain-specifically versus globally phrased scales. Additionally, when employing grades from different subjects as criteria and comparing the predictive power of all variance sources in the model, the mechanisms leading to the predictive power of each measure were hypothesized to become clearer.

Method

Sample

Three hundred twenty-five school children (174 females, 151 males) participated in the present study. They attended two different schools and school types: 174 of them German “Hauptschule” (secondary school, which offers Lower Secondary Education according to the International Standard Classification of Education; ISCED) and 151 German “Gymnasium”

(secondary school, which prepares students to enter university) in either 8th (n=194) or 9th (n=131) grade. Ages ranged from 13 to 17 years, with an average of 14.32 years (SD=0.92).

Procedure

All the data assessed in the present study were based on student self-reports via questionnaires (for a description of these questionnaires, see the Instruments section). The assessments took place during regular class hours, and participation was voluntary for the students. Within the questionnaire, students first rated their global performance approach (6 items) and avoidance (6 items), mastery approach (5 items) and avoidance (3 items), hope for success (5 items) and fear of failure (5 items) tendencies on a 4-point rating scale ranging from 1 (*not at all*) to 4 (*very much*). In the second part of the questionnaire, students had to rate their personality using the same 4-point rating scale. In the third part of the questionnaire, students answered the same questions as before regarding performance, mastery, approach, avoidance, hope for success, and fear of failure, only this time the items were focused specifically on the domain of mathematics.

In the last section of the questionnaire, students reported their grades in math, German, and physics. In the German education system, grades range from 1, the very best, to 6, the worst grade, with 5 and 6 indicating insufficient performance. Whereas there might be concern about the use of self-report grades, Kuncel, Crede, and Thomas (2005) showed that such self-reported grades provide a valid representation of students' actual academic performance. Similar results were reported by Dickhäuser and Plenter (2005) for a German sample. Math grades were lowest (M=3.11, SD=1.01), followed by grades in German language (M=2.95, SD= 0.73) and physics (M=2.83, SD= 0.93). The zero-order correlations between the grades were r=.51 for Math and Physics, r = .31 for Math and German, and r = .24 for Physics and German (with $p < .01$ for all correlations). Grades for math and physics showed very similar distributions, with German grades being narrower and less differentiating.

Instruments

The approach versus avoidance by mastery versus performance motivation scales were from Elliot and McGregor's Achievement Goal Questionnaire (2001).The original AGQ is phrased in a specific manner because the items include markers such as "in this class" (e.g., "My goal in this class is to get a better grade than the other students").Therefore, for the globally phrased part of the questionnaire, we adapted these items slightly and used a general phrase: "In general, it is my goal to get a better grade than the other students." Conversely, for the math-domain-specific items, we changed the item to, for example: "My goal in math class is to get a better grade than the other students."

The revised German Achievement Motives Scale (AMS-R; Lang & Fries, 2006; cf. Nygård & Gjesme, 1973) consists of the scales hope for success and fear of failure. Items from the AMS-R read "I like situations in which I can find out how capable I am" or "I am afraid of failing in somewhat difficult situations when a lot depends on me." Again, for the globally phrased part of the questionnaire, we had to remove any hints at frames of references (e.g., "I like it when I can find out how capable I am").By contrast, for the domain-specifically phrased part, these items were altered slightly to reflect the intention to focus solely on mathematics. This was mostly done by just adding "in math" or "in math class," by deleting adverbs denoting restrictions like "somewhat," and by linking the assertion to math. For example, the item mentioned above would read as follows in the math-specific version: "In

math, I like it when I can find out how capable I am."Similar modifications to the AGQ to reflect a domain-specific approach have been done before (e.g., Finney, et al., 2004) where items have been rewritten in reference to achievement during the current semester but not for a specific class as was done here. Finney et al. (2004) showed that these modifications did not alter the factorial validity.

The BFI-K (Rammstedt & John, 2005) consists of 21 items, with 4 items for each domain of the Big Five except for Openness, which is assessed with 5 items. Although the standard answer format for the BFI-K is a 5-point rating scale, we opted to use the same 4-point rating scale as was used for the other scales in the current study in order to avoid confusing the students. By doing so, the hard-to-interpret midpoint of the scale was eliminated (Kulas & Stachowski, 2009; Rost, Carstensen, & von Davier, 1999).

Because the answer format of the BFI-K had been changed in this study, scores are not comparable to the numbers given in the original publication. However, the same pattern regarding reliabilities and means as described in the original publication emerged in our sample: Internal consistency was lowest for Agreeableness (α=.45) and Openness (α =.55), and highest for Extraversion (α =.70), with Neuroticism (α =.68) and Conscientiousness (α =.58) in between.

Internal consistencies and descriptive statistics for most motivational scales were satisfactory with Cronbach's alphas typically ranging above .80. Exceptions were the mastery motivation scales framed generally, which had comparatively low Cronbach's alphas of .48 and .44 for mastery approach and avoidance, respectively. All Cronbach's alpha statistics of the scales employed are presented in Table 1.

Statistical Analyses

Statistical tests were conducted using PASW™ 18 and G*Power 3.1.2 (Faul, Erdfelder, Lang, & Buchner, 2007), and the structural equation models (SEM) were analyzed in Mplus™ 5.2 (Muthén & Muthén, 1998-2007) using robust maximum likelihood estimation (MLR) (a) to take into account that student ratings were nested within two different schools, and (b) in order to receive significant values that were robust against a violation of multivariate normality. Besides the χ^2test, assessment of the global goodness-of-fit was based on the Standardized Root Mean Square Residual (SRMR) and the Root Mean Square Error of Approximation (RMSEA) as recommended by Hu and Bentler (1999), and on the Comparative Fit Index (CFI) as advised by Beauducel and Wittmann (2005). Following Hu and Bentler's advice, we used the following cutoff criteria for assuming good model fit: the SRMR should be smaller than .11, the RMSEA should be less than or equal to .06, and the CFI should have a value of approximately .95.

Models

An MTMM structural equation model was specified to separate the variance from the components inside the various measures (see Figure 1). In an attempt to disentangle the different variances of each measure, we had loadings from the appropriate measures on six latent factors representing the different approaches to the need for achievement adopted in the questionnaires used; namely, approach, avoidance, mastery, performance, hope for success, and fear of failure. Additionally, we had loadings from all of the globally phrased measures on one latent "globality" factor and loadings from all of the math-specifically phrased

measures on one "domain" factor. To test our hypotheses, we used three different models. First, we had to establish the fit of the basic model as depicted in Figure 1.

The second model (criterion Models 2a-c) built upon the first model. This model additionally incorporated school grades. For that, we correlated each of the eight latent factors with either math or physics or German school grades to estimate the size of the relationship of each factor with the grades.

The third model (criterion Models3a-c), again, built upon the second model. This time, each school grade was regressed onto the personality domain conscientiousness. The resulting correlations of the eight latent factors with the school grades thus were semi-partial correlations adjusted for the influence of conscientiousness.

Results

Comparison of Reliabilities and Means

As can be gathered from the main diagonal values in Table 1 representing the reliabilities of the scales, the specifically phrased scales universally had better internal consistency. Using the Feldt test to compare alpha coefficient sizes (Alsawalmeh & Feldt, 1994), the global-specific comparison for each of the six scales was significant at $p < .001$. This shows that the domain-specific measures had better internal consistencies.

Zero-order intercorrelations of the domain-specifically phrased scales among each other were higher than the equivalent correlations among the globally phrased scales (values of differences ranged from Studentized $z = 0.57$ to $z = 4.24$[2]). This was true for all measures except for fear of failure, which showed a trend toward a smaller correlation when phrased specifically. Furthermore, all correlations of fear of failure with the other constructs were the lowest. This shows that the domain-specific measures share more common systematic variance.

Finally, we compare mean values of the scales for the globally versus specifically phrased versions: Generally, the means tended to be lower in the domain-specific than in the global versions, with the exception of mastery avoidance and hope for success where there was virtually no difference at all. Apparently, framing the scales in a domain-specific manner induced participants to differentiate more between the various aspects of motivation and to endorse smaller values.

Predictive Validity of the Globally versus Specifically Phrased Scales

Table 2 presents the test-criterion correlations of the scales with grades in math, physics, and German. As can be seen, for math, all specific scales attained moderate and significant test-criterion correlations with the exception of mastery avoidance. Moreover, the domain-specifically phrased scales produced systematically higher correlations than the globally phrased scales (z-values from $z = 0.51$ to $z = 2.85$). For the prediction of physics grades, the results were different: Both kinds of scales produced test-criterion correlations that were comparable in their size. As could be expected, this was mainly due to a reduction in test-

[2] Values $\geq |1.96|$ denote significance at $p \leq .05$.

criterion correlations for the specifically phrased scales. This means the correlations with the physics grade were not only comparable between both kinds of scales but also comparable with the test-criterion correlation of the global scale with math. In other words, for a grade from an adjacent subject, the domain-specifically phrased scales worked no better than the globally phrased scales (z-values for the differences ranged from $z < 0.001$ to $z = 0.522$). Consequently, the same four scales reached small and significant correlations for both the global and the specific phrasings with physics grades. For German grades, all scales yielded the lowest test-criterion correlations of the three subjects, only two of which achieved significance. These exceptions were mastery avoidance (specific) and hope for success (global). Differences between global and specific measures ranged from $z = 0.38$ to $z = 1.28$.

Taken together, the scales phrased specifically for math provided the best predictions when used to predict domain-congruent grades (i.e., math). For the other, non-domain-congruent grades, it did not make any difference whether globally or specifically phrased scales were used; that is, the domain-specific scales worked similarly to the global scales once the domains did not match.

Latent SEM Approach: Modeling the Components of Variance as Latent Variables

Basic MTMM-model with confirmatory factor analyses (CFA). Model fit and fit indices for the basic model (Figure 1) can be found in Table 3. All models converged properly. The basic model achieved an excellent fit. Models 2a-c, with the eight correlations from the latent factors to the three different grades, achieved excellent fit as well. Models3a-c, with the regression of grades onto conscientiousness, achieved acceptable fit, being otherwise similar again across the three grades. The worsening of the fit can be traced back to the newly introduced variable of conscientiousness not being allowed to correlate with any of the other latent factors or their indicators.

Comparison of communalities and weights for globally versus specifically phrased scales. Table 4 gives the communalities (h^2) of the scales and their standardized weights on the latent variables as specified in the basic model (see Figure 1). Communalities for both the specifically and the globally phrased scales mostly mirrored the differences observed for the reliabilities, with slightly higher values for the specifically phrased versions. The exceptions were mastery avoidance and fear of failure. The loading patterns on the motivational latent trait variables, on the other hand, were comparable for the globally and the domain-specifically phrased scales, both when comparing loading patterns within a latent variable (i.e., vertically) and when comparing patterns across latent variables (i.e., horizontally). Again, there was the exception of mastery avoidance. Whereas the globally phrased version loaded highly on the latent mastery variable, the loading from the specifically phrased version was quite a bit lower (global: .63 vs. specific: .38); conversely, the loading on the latent avoidance variable was inverted, with a higher loading from the specifically phrased version (specific: .44 vs. global: .26). When looking at the latent method variables “global” and “domain,” which should capture the variance due to different phrasing, loadings from the specifically phrased scales, overall, were a bit higher than those from the globally phrased scales except for the two performance scales and the mastery avoidance scale.

Analysis of correlation patterns without controlling for conscientiousness. Each left column of Table 5 presents the zero-order correlations and respective significances of the eight latent variables in the model with the three grades. For all of the three grades, the latent method variable “domain” produced significant correlations. This correlation was highest for

math, decreased for physics, and even reversed its sign for German. Aside from this correlation, with respect to math grades, only fear of failure and mastery produced significant correlations: The more fear someone experienced, the worse their math grade; the more someone aimed for mastery, the better their math grade. For physics grades, fear of failure and mastery exhibited the same correlations as for math grades. Additionally, the latent variable approach was significantly related to physics grades. For German grades, a pattern different from the two natural science topics turned up: The correlations for the latent variables performance and avoidance became significant, whereas fear of failure and mastery were not significant anymore. Thus, the more someone aimed for performance goals and tried to avoid failure, the better their German grade. Generally, the correlations with German grades were the lowest of the three subjects.

Analysis of correlation patterns with grades adjusted for conscientiousness. Each right column of Table 5 presents the semi-partial correlations of the eight latent variables in the models with the three grades after the grades had been adjusted for conscientiousness (the regression weights for conscientiousness were $\beta = -.12$ for math, $\beta = -.15$ for physics, and $\beta = -.14$ for German, with all three ps $<.001$). Even after controlling for conscientiousness, the overall pattern of correlations remained the same for all three grades. Accordingly, the test-criterion correlation of the latent variable capturing the "domain" variance retained its size, direction, and significance for the three grades. Furthermore, the same patterns as before for fear of failure and mastery re-emerged: Both correlations reached significance for math grades and physics grades, but not so for German grades. There are some noteworthy differences when comparing to the model without conscientiousness: Two of the correlations missed significance: (a) the correlation of approach with physics grades and (b) the correlation of avoidance with German grades.

Discussion

The present study aimed at elucidating possible explanations for the higher test-criterion correlations for measures phrased in a domain-specific fashion. To this end, we used achievement motivation measures (a) phrased globally versus (b) phrased specifically for the domain of math as predictors of math performance at school. Results confirmed the higher test-criterion correlations for the domain-specific relative to the generally phrased achievement motivation measures. Furthermore, whereas reliabilities and intercorrelations for the specific measures were higher, their means were lower compared with the global measures. MTMM analyses revealed that all measures had comparable trait satiation, with the specific measures having higher loadings on their method factor (Aiken & Groth-Marnat, 2006). The disentangling of the variance sources further showed that test-criterion correlations for the domain-specifically phrased scales were mainly due to the specific variance. This held true even after controlling for conscientiousness. Below, these results will be outlined in more detail, and two possible explanations for the higher test-criterion correlations of the domain-specific measures will be discussed: Specific measures capture either (a) a narrower facet such as math motivation or (b) an additional construct.

Means, Reliabilities, and Intercorrelations

When looking at the means for the globally and domain-specifically phrased scales, it was apparent that all of the specific scales had lower means compared with the respective global scales. Apparently, when participants are asked to rate their motivation for math, they end up with lower self-evaluations than when asked about their global motivation. This could be

because math is generally perceived as very demanding and therefore seems less surmountable using only dedication and commitment. This higher demand from or more careful evaluation of math can be gathered also from the generally lower grades in this subject compared to other subjects that occurred in our data (Sabot & Wakeman-Linn, 1991). Another explanation could be that adding a specific subject to the item yields more careful self-evaluations. Nonetheless, the differences in means suggest that something different is being assessed. At the core of this could be either a narrower trait motivation (i.e., math motivation) or another construct besides achievement motivation. However, it is hard to imagine that the assessment of a narrower trait would result in uniformly lower scores regardless of the emotional valence of the scales. Specifically, a specific math motivation should mostly affect the negatively valenced scales because math is, as was mentioned above, the subject with the most negative associations according to students. On the other hand, an additional source of variance could manifest itself equally in all scales.

The higher reliabilities of all of the domain-specifically phrased scales show that they contain more systematic variance. This gain in systematic variance may originate in the items being more homogeneous (i.e., measuring a narrower construct range). However, empirical evidence in other research has usually shown that reliabilities on the facet level are inferior to the reliabilities of scales that assess higher-level constructs (Costa & McCrae, 1992). Thus, this explanation seems rather unlikely. Another explanation could be that the specific formulation of the items introduces a further systematic source of variance. This would mean that the items no longer capture only one of the need-for-achievement-related goals or motives they were designed to capture, but capture an additional construct as well. Because higher reliabilities can be found uniformly for all scales regardless of their valence (negative or positive) or their time frame (prior, during, or after the achievement situation, depending on the theoretical school), this additional source of variance very likely is related to the math-specific phrasing and not the achievement-related item content. Thus, we suggest that an additional construct is being assessed. This is similar to a spurious measurement error, which stems from a systematic second source of variance (e.g., Ziegler & Buehner, 2009).

Regarding the intercorrelations of the differently phrased scales, it has to be kept in mind that all scales but FF and avoidance are positively valenced. Because of the rather negative valence of math as a school subject, only the negatively valenced scales would exhibit higher intercorrelations if a narrower trait motivation were being measured. But that was not the case; instead, all intercorrelations of the specific scales were higher. This makes it quite plausible that another construct as an additional source of variance other than achievement motivation is at the core of these higher correlations.

MTMM Analyses

Loadings and communalities. Trait loadings were found to be equal for all scales regardless of their phrasing. This means that regardless of whether a scale is phrased globally or domain-specifically, the amount of systematic variance that is due to the achievement-motivation-related construct measured is comparable. However, the amount of variance explained by the respective method factor was larger for domain-specific scales. In other words, the phrasing of items with specific reference to the domain of mathematics introduces new and systematic variance that cannot be explained by the general achievement motivation constructs. Again, this could be explained by either a narrower trait whose variance ends up in the “domain” variable (i.e., the method factor) or an additional source of variance. The fact that all domain-specific scales have higher method factor loadings can be explained only by

assuming that all these scales now have more common variance than the global scales. These higher loadings can be found regardless of whether a scale has positive or negative valence. This would mean that whatever variance is math specific in these scales would be the same in each trait. This again seems to speak against specific narrower facets subsumed under each of the global constructs, but rather for an additional construct that is being tapped.

Test criterion correlations. ***Traits as predictors.*** Our data revealed that when controlling for method variance, test-criterion correlations of the trait variables were generally low regardless of the school subject used as criterion. The only exception to this seems to be FF: only the negatively valenced FF retained its test-criterion validity, and only for math and physics but not for German. This different pattern across subjects may be explained by the fact that the native language is on the other end of the "math-verbal" line of school subjects (Marsh, Byrne, & Shavelson, 1988; Marsh & Yeung, 1996). It is conceivable that the subject of German is less anxiety-driven than the subjects situated on the math-pole (Marsh, 1988) because "everybody" is confident about knowing her or his native language (Achen & Courant, 2009).

These results do not change when controlling for conscientiousness. The β-weights for conscientiousness were quite similar for all three grades. For German grades, conscientiousness was the only trait having a small and significant influence on grades. This could mean that when considering the measures used in this study, grades in German mostly depend on what work attitude someone exhibits as captured by the conscientiousness scale; at least more so than grades in math and physics. Still, the influence of the trait variables on grades was low and mostly independent of the school subject.

Method factors as predictors.The latent method variable "domain" showed the highest test-criterion correlation of all variance sources when using grades in the specific subject (i.e., math) as a criterion. This moderate correlation fell to a smaller size for grades in the adjacent subject (i.e., physics), and reversed its sign for German. This S-shaped pattern makes it highly implausible that the variance of this "domain" method variable is but a pure method-artifact because, if so, it should not have any substantial correlations with other grades except the matching one. Interestingly, the same pattern of influence was already found in self-concept research: The mathematics self-concept was found to be negatively related to performance in the verbal domain, which was explained by the "internal/external frame of reference model." According to this model, the subjective evaluation of one's performance is based on not only social comparisons (external frame of reference), but also on cross-domain or dimensional comparisons (internal frame of reference) implying a comparison of one's own achievement in a subject area with achievement in other subjects (Marsh, 1986, 1990; Möller & Köller, 2004; Schilling, Sparfeldt, Rost, & Nickels, 2005). The similarity of the correlative pattern of our method factor "domain" with mathematics/physics and German grades and the correlative pattern of mathematics self-concept and the same grades clearly supports the earlier interpretations that an additional source of variance is being captured in the specifically phrased scales. Furthermore, it points out a possible candidate for this source: the self-concept. Apparently, self-concepts in school are shaped by getting specific feedback on achievement, which consists mainly of grades (D. H. Rost, Sparfeldt, Dickhauser, & Schilling, 2005). Accordingly, the self-concept is thought to originate in performance (Marsh, 1992). Bandura (1997) argued that a self-concept can influence performance more than actual knowledge or ability in a certain domain. In addition, it was already reported by Marsh (1986) that a positive verbal self-concept has negative effects on math-achievement, and a

positive math self-concept has a negative effect on verbal-achievement. This reflects the test-criterion correlation found for the "domain" method factor.

The results found here clearly show that more systematic variance was captured by the scales that were phrased in a domain-specific fashion. Given our findings, which showed that the differences between global and domain-specific were similar across motivational constructs, this being caused by a narrower trait seems unlikely. The alternative explanation that an additional construct is being assessed is much more in line with these results. As already hinted at above, this could very well be self-concept. However, the only way to conclusively answer the question of whether it is really self-concept inside the "domain" factor, a self-concept questionnaire specific for each different school subject should be employed alongside the motivational measures as a convergent-validity-related measure.

Limitations

One limiting factor in this study was the use of only math-specific scales. This limited the power of possible explanations for the additional variance found in the domain-specific scales. Future studies should incorporate specific measures for different school subjects as well as specific self-concept measures. Such a triangulation would further shed light on the sources of variance found in these domain-specific scales through these divergent measures.

Another limiting factor was the use of only 8thand 9thgraders. Future studies should include older students and university students to corroborate our findings.

Additionally, our findings could be specific to the scales used in this study. More and different measures of the need for achievement should be employed in future research.

Conclusion

The present results suggest that we are capturing more systematic variance when employing specific measures of motivation, and that we can make better predictions in doing so. This increase in explained variance, however, seems to be due not to these so-called specific measures providing better measures of "motivation," but rather to the concept of one's self from which a person draws her or his conclusions regarding her or his commitment. That this self-concept is not equivalent to global trait motivation and probably also not equivalent to specific trait motivation can be demonstrated by comparing these global and specific measures directly and also by employing an MTMM-SEM approach. Thus, whenever test-criterion correlations for specific measures are reported, conclusions regarding the actual mechanisms should include the possibility that self-concept is the driving force behind the correlation.

References

Achen, A. C., & Courant, P. N. (2009). What Are Grades Made Of? Journal of Economic Perspectives*, 23*, 77-92.

Aiken, L. R. (1976). Update on attitudes and other affective variables in learning mathematics. Review of Educational Research*, 46*, 293.

Aiken, L. R., & Groth-Marnat, G. (2006). Psychological testing and assessment. Boston: Pearson Education Group.

Alsawalmeh, Y. M., & Feldt, L. S. (1994). A Modification of Feldt Test of the Equality of 2 Dependent Alpha-Coefficients. Psychometrika*, 59*, 49-57.

Atkinson, J. W. (1957). Motivational Determinants of Risk-Taking Behavior. Psychological Review*, 64*, 359-372.

Bagby, R. M., Costa, P. T., Widiger, T. A., Ryder, A. G., & Marshall, M. (2005). DSM-IV personality disorders and the five-factor model of personality: A multi-method examination of domain- and facet-level predictions. European Journal of Personality*, 19*, 307-324.

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. Structural Equation Modeling*, 12*, 41-75.

Bong, M. (2001). Between- and within-domain relations of academic motivation among middle and high school students: Self-efficacy, task-value, and achievement goals. Journal of Educational Psychology*, 93*, 23-34.

Bouffard, T., Boisvert, J., Vezeau, C., & Larouche, C. (1995). The Impact of Goal Orientation on Self-Regulation and Performance among College-Students. British Journal of Educational Psychology*, 65*, 317-329.

Bouffard, T., Vezeau, C., Romano, G., Chouinard, R., Bordeleau, L., & Filion, C. (1998). Élaboration et validation d'un instrument pour évaluer les buts des élèves en contexte scolaire [Development and validation of a questionnaire on goals in academic settings]. Revue Canadienne Des Sciences Du Comportement [Canadian Journal of Behavioural Science]*, 30*, 203-206.

Brunswik, E. (1955). Representative Design and Probabilistic Theory in a Functional Psychology. Psychological Review*, 62*, 193-217.

Campbell, F. A., & Ramey, C. T. (1994). Effects of Early Intervention on Intellectual and Academic-Achievement - a Follow-up-Study of Children from Low-Income Families. Child Development*, 65*, 684-698.

Ciani, K. D., & Sheldon, K. M. (2010). Evaluating the mastery-avoidance goal construct: A study of elite college baseball players. Psychology of Sport and Exercise*, 11*, 127-132.

Clark, R. A., Teevan, R., & Ricciuti, H. N. (1956). Hope of Success and Fear of Failure as Aspects of Need for Achievement. Journal of abnormal and social psychology, *53*, 182-186.

Conroy, D. E. (2003). Representational models associated with fear of failure in adolescents and young adults. Journal of Personality, *71*, 757-783.

Costa, P. T., & McCrae, R. R. (1992). Revised NEO personality inventory (NEO-PI-R) and NEO Five-Factor inventory (NEO-FFI): Professional Manual.: Odessa: Psychological Assessment Resources.

Covington, M. V. (2000). Goal theory, motivation, and school achievement: An integrative review. Annual Review of Psychology, *51*, 171-200.

Dickhäuser, O., & Plenter, I. (2005). Letztes Halbjahr stand ich zwei. Zur Akkuratheit selbst berichteter Noten [On the accuracy of self-reported school marks]. Zeitschrift für Pädagogische Psychologie, *19*, 219-224.

Dweck, C. S. (1986). Motivational processes affecting learning. American Psychologist, *41*, 1040-1048.

Dweck, C. S., & Leggett, E. L. (1988). A Social Cognitive Approach to Motivation and Personality. Psychological Review, *95*, 256-273.

Elliot, A. J., & Harackiewicz, J. M. (1996). Approach and avoidance achievement goals and intrinsic motivation: A mediational analysis. Journal of Personality and Social Psychology, *70*, 461-475.

Elliot, A. J., & McGregor, H. A. (2001). A 2 x 2 achievement goal framework. Journal of Personality and Social Psychology, *80*, 501-519.

Elliot, A. J., & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. Journal of Educational Psychology, *100*, 613-628.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, *39*, 175-191.

Finney, S. J., Pieper, S. L., & Barron, K. E. (2004). Examining the psychometric properties of the Achievement Goal Questionnaire in a general academic context. Educational and Psychological Measurement, *64*, 365-382.

Fleeson, W., & Noftle, E. E. (2008). Where Does Personality Have Its Influence? A Supermatrix of Consistency Concepts. Journal of Personality, *76*, 1355-1386.

Gjesme, T. (1981). Is there any future in achievement motivation? Motivation and Emotion, *5*, 115-138.

Gjesme, T., & Nygård, R. (1970). Achievement-related motives: Theoretical considerations and construction of a measuring instrument. Unpublished report. University of Oslo.

Goetz, T., Frenzel, A. C., Pekrun, R., & Hall, N. C. (2006). The domain specificity of academic emotional experiences. Journal of Experimental Education, *75*, 5-29.

Goetz, T., Frenzel, A. C., Pekrun, R., Hall, N. C., & Lüdtke, O. (2007). Between- and within-domain relations of students' academic emotions. Journal of Educational Psychology, *99*, 715-733.

Green, J., Martin, A. J., & Marsh, H. W. (2007). Motivation and engagement in English, mathematics and science high school subjects: Towards an understanding of multidimensional domain specificity. Learning and Individual Differences, *17*, 269-279.

Greene, B. A., Miller, R. B., Crowson, H. M., Duke, B. L., & Akey, K. L. (2004). Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation. Contemporary Educational Psychology, *29*, 462-482.

Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002). Revision of achievement goal theory: Necessary and illuminating. Journal of Educational Psychology, *94*, 638-645.

Heckhausen, H. (1977). Achievement Motivation and Its Constructs: A Cognitive Model. Motivation and Emotion, *1*, 283-329.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, *6*, 1-55.

Ironson, G. H., Brannick, M. T., Smith, P. C., Gibson, W. M., & Paul, K. B. (1989). Construction of a Job in General Scale - a Comparison of Global, Composite, and Specific Measures. Journal of Applied Psychology, *74*, 193-200.

Kulas, J. T., & Stachowski, A. A. (2009). Middle category endorsement in odd-numbered Likert response scales: Associated item characteristics, cognitive demands, and preferred meanings. Journal of Research in Personality, *43*, 489-493.

Kuncel, N. R., Crede, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. Review of Educational Research, *75*, 63-82.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. Psychological Bulletin, *127*, 162-181.

Lang, J. W. B., & Fries, S. (2006). A revised 10-item version of the Achievement Motives Scale. Psychometric properties in German-speaking samples; Eine revidierte 10-Item-Version der Achievement Motives Scale (Skala zu Leistungsmotiven). European Journal of Psychological Assessment, *22*, 216-224.

Linnenbrink, E. A., & Pintrich, P. R. (2002). Achievement goal theory and affect: An asymmetrical bidirectional model. Educational Psychologist, *37*, 69-78.

Ma, X. (1999). A Meta-Analysis of the Relationship between Anxiety toward Mathematics and Achievement in Mathematics. Journal for Research in Mathematics Education, *30*, 520-540.

MacCann, C., Duckworth, A., & Roberts, R. (2009). Empirical identification of the major facets of Conscientiousness. Learning and Individual Differences, *19*, 451-458.

Marsh, H. W. (1986). Verbal and Math Self-Concepts – an Internal External Frame of Reference Model. American Educational Research Journal, *23*, 129-149.

Marsh, H. W. (1988). The content specificity of math and english anxieties: The high school and beyond study. Anxiety Research, *1*, 137-149.

Marsh, H. W. (1990). Influences of Internal and External Frames of Reference on the Formation of Math and English Self-Concepts. Journal of Educational Psychology, *82*, 107-116.

Marsh, H. W. (1992). Content Specificity of Relations between Academic-Achievement and Academic Self-Concept. Journal of Educational Psychology, *84*, 35-42.

Marsh, H. W. (1993). The Multidimensional Structure of Academic Self-Concept - Invariance over Gender and Age. American Educational Research Journal, *30*, 841-860.

Marsh, H. W., Byrne, B. M., & Shavelson, R. J. (1988). A Multifaceted Academic Self-Concept - Its Hierarchical Structure and Its Relation to Academic-Achievement. Journal of Educational Psychology, *80*, 366-380.

Marsh, H. W., & Yeung, A. S. (1996). The distinctiveness of affects in specific school subjects: An application of confirmatory factor analysis with the National Educational Longitudinal Study of 1988. American Educational Research Journal, *33*, 665-689.

Martin, A. J. (2008). How domain specific is motivation and engagement across school, sport, and music? A substantive-methodological synergy assessing young sportspeople and musicians. Contemporary Educational Psychology, *33*, 785-813.

McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). The achievement motive. New York: Appleton-Century-Crofts.

Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. Journal of Educational Psychology, *82*, 60-70.

Möller, J., & Köller, O. (2004). Die Genese akademischer Selbstkonzepte: Effekte dimensionaler und sozialer Vergleiche [On the development of academic self-concepts: The impact of social and dimensional comparisons]. Psychologische Rundschau, *55*, 19-27.

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

Muthén, L. K., & Muthén, B. (1998-2007). Mplus User's Guide. Fifth Edition. Los Angeles, CA: Muthén & Muthén.

Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. Psychological Review*, 91*, 328-346.

Nygård, R., & Gjesme, T. (1973). Assessment of Achievement Motives: Comments and Suggestions. Scandinavian Journal of Educational Research*, 17*, 39-46.

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. Personality and Individual Differences*, 43*, 971-990.

Paunonen, S. V., & Ashton, M. C. (2001). Big five factors and facets and the prediction of behavior. Journal of Personality and Social Psychology*, 81*, 524-539.

Perkins, D. N., & Grotzer, T. A. (1997). Teaching intelligence. American Psychologist*, 52*, 1125-1133.

Poropat, A. E. (2009). A Meta-Analysis of the Five-Factor Model of Personality and Academic Performance. Psychological Bulletin*, 135*, 322-338.

Rammstedt, B., & John, O. P. (2005). Short version of the Big Five Inventory (BFI-K): Development and validation of an economic inventory for assessment of the five factors of personality. Diagnostica*, 51*, 195-206.

Roberts, B. W., Chernyshenko, O. S., Stark, S., & Goldberg, L. R. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. Personnel Psychology*, 59*, 103-139.

Rosenberg, M., Schooler, C., Schoenbach, C., & Rosenberg, F. (1995). Global Self-Esteem and Specific Self-Esteem - Different Concepts, Different Outcomes. American Sociological Review*, 60*, 141-156.

Rost, J., Carstensen, C. H., & von Davier, M. (1999). Are the Big Five Rasch scalable? A reanalysis of the NEO-FFI norm data. Diagnostica*, 45*, 119-127.

Sabot, R., & Wakeman-Linn, J. (1991). Grade Inflation and Course Choice. The Journal of Economic Perspectives*, 5*, 159-170.

Schilling, S. R., Sparfeldt, J. R., Rost, D. H., & Nickels, G. (2005). Facets of academic self-concept - Validity of the Differential Self-Concept Grid (DISC-Grid). Diagnostica*, 51*, 21-28.

Schunk, D. H. (1991). Self-Efficacy and Academic Motivation. Educational Psychologist*, 26*, 207-231.

Steinmayr, R., & Spinath, B. (2007). Predicting school achievement from motivation and personality. Zeitschrift fur Pädagogische Psychologie*, 21*, 207-216.

Steinmayr, R., & Spinath, B. (2009). The importance of motivation as a predictor of school achievement. Learning and Individual Differences*, 19*, 80-90.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. Psychological Review*, 92*, 548-573.

Wigfield, A. (1997). Reading motivation: A domain-specific approach to motivation. Educational Psychologist*, 32*, 59-68.

Wigfield, A., Guthrie, J. T., Tonks, S., & Perencevich, K. C. (2004). Children's motivation for reading: Domain specificity and instructional influences. Journal of Educational Research*, 97*, 299-309.

Ziegler, M., & Buehner, M. (2009). Modeling Socially Desirable Responding and Its Effects. Educational and Psychological Measurement*, 69*, 548-565.

Ziegler, M., Danay, E., Schölmerich, F., & Bühner, M. (2010). Predicting Academic Success with the Big 5 Rated from Different Points of View: Self-Rated, Other Rated and Faked. European Journal of Personality*, 24*, 341-355.

Ziegler, M., Knogler, M., & Bühner, M. (2009). Conscientiousness, Achievement Striving, and Intelligence as Performance Predictors in a Sample of German Psychology Students: Always a Linear Relationship? Learning and Individual Differences*, 19*, 288-292.

Ziegler, M., Schmukle, S., Egloff, B., & Bühner, M. (2010). Investigating Measures of Achievement Motivation(s). Journal of Individual Differences*, 31*, 15-21.

Table 1
Descriptive Statistics and Independent t Tests Comparing Gobal and Domain-Specific Motivational Measures)

					Global						*Domain-specific*					
					P_{Ap}	P_{Av}	M_{Ap}	M_{Av}	HS	FF	P_{Ap}	P_{Av}	M_{Ap}	M_{Av}	HS	FF
		M	*SD*	*t* [d_Z] (df=324)	*r*						*r*					
Global	P_{Ap}	1.68	.61	5.28^{***} [.29]	**.78**											
	P_{Av}	1.68	.67	4.21^{***} [.23]	$.66^{**}$	**.77**										
	M_{Ap}	2.00	.56	2.24^{*} [.12]	$.35^{**}$	$.20^{**}$	**.48**									
	M_{Av}	1.58	.61	-.25 [.01]	$.23^{**}$	$.18^{**}$	$.31^{**}$	**.44**								
	HS	1.94	.58	.35 [.02]	$.28^{**}$	$.14^{*}$	$.40^{**}$	$.15^{**}$	**.67**							
	FF	1.27	.63	5.92^{***} [.33]	.11	$.28^{**}$	$.18^{**}$	.11	.06	**.71**						
Domain	P_{Ap}	1.56	.69		$.79^{**}$	$.57^{**}$	$.35^{**}$	$.21^{**}$	$.26^{**}$	.05	**.85**					
	P_{Av}	1.56	.75		$.66^{**}$	$.72^{**}$	$.24^{**}$	$.19^{**}$	$.12^{*}$	$.19^{**}$	$.74^{**}$	**.85**				
	M_{Ap}	1.93	.69		$.36^{**}$	$.23^{**}$	$.58^{**}$	$.28^{**}$	$.40^{**}$	.11	$.56^{**}$	$.38^{**}$	**.84**			
	M_{Av}	1.59	.67		$.18^{**}$	$.23^{**}$	$.31^{**}$	$.35^{**}$	$.12^{*}$	$.19^{**}$	$.29^{**}$	$.34^{**}$	$.35^{**}$	**.57**		
	HS	1.93	.64		$.29^{**}$	$.23^{**}$	$.33^{**}$	$.16^{**}$	$.55^{**}$	.08	$.46^{**}$	$.34^{**}$	$.64^{**}$	$.26^{**}$	**.80**	
	FF	1.09	.74		.03	$.15^{**}$	$.24^{**}$	.09	.05	$.71^{**}$	-.03	$.11^{*}$	.06	$.19^{**}$	-.04	**.84**

Note. *N*=325; Internal consistencies (α) are bold numbers in diagonal. P=performance; M=mastery; HS=hope for success; FF=fear of failure; Ap=approach; Av=avoidance.
$^{*}p$ <.05. $^{**}p$< .01. $^{***}p$< .001.

Table 2
Correlations between Different Measures of Achievement Motivation and Different Grades

		Math grades *r*	Physics grades *r*	German grades *r*
Global	Performance approach	-.16**	-.13*	-.04
	Performance avoidance	-.07	-.08	-.02
	Mastery approach	-.10	-.15**	-.06
	Mastery avoidance	-.06	-.07	<.001
	Hope for success	-.14**	-.15**	-.12
	Fear of failure	.18**	.12*	<.01
Domain	Performance approach	-.30***	-.13*	-.03
	Performance avoidance	-.18**	-.09	-.05
	Mastery approach	-.31***	-.19**	.03
	Mastery avoidance	-.10	-.09	-.08
	Hope for success	-.35***	-.15**	-.02
	Fear of failure	.31***	.15**	.02

Note. N = 325.
*$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

Table 3
Model-Fit and Fit Indices for the Three Different Models, Divided by Correlated Grades

Model	$\chi^2(df)$	*p*	RMSEA (90% CI)	CFI	SRMR
MTMM (basic) (ap x av; ap x ma)	24.81(30)	.73	.000 (.000-.032)	1.0	.031
MTMM math	27.52(34)	.78	.000 (.000-.028)	1.0	.030
MTMM physics	29.78(34)	.67	.000 (.000-.033)	1.0	.030
MTMM German	33.36(34)	.50	.000 (.000-.039)	1.0	.030
MTMM math + conscientiousness	100.01(46)	<.001	.060 (.044-.076)	.981	.084
MTMM physics + conscientiousness	106.75(46)	<.001	.064 (.048-.080)	.980	.084
MTMM German + conscientiousness	113.88(46)	<.001	.068 (.052-.083)	.979	.084

Note. Estimator: MLR for clustered data. $\chi^2(df)$= χ^2 value and degrees of freedom; RMSEA (90% CI) = root mean square error of approximation with 90% confidence interval; CFI = comparative fit index; SRMR = standardized root mean square residual.

Table 4
Standardized Regression Weights of the Latent Variables on the Different Scales, and h² Values of These Scales

	Perform.	Mastery	Approach	Avoid.	HS	FF	global	domain	h^2
P_{Ap} (*g*)	.79***		.46***				.23***		.88***
P_{Av} (*g*)	.60***			.52***			.36**		.75***
M_{Ap} (*g*)		.03	.77***				-.006		.60***
M_{Av} (*g*)		.63***		.26***			.12		.47***
HS (*g*)			.51***		.59***		.005		.61***
FF (*g*)				.38***		.80***	.29**		.87***
P_{Ap} (*s*)	.72***		.49***					.36***	.88***
P_{Av} (*s*)	.68***			.58***				.29***	.88***
M_{Ap} (*s*)		-.03	.78***					.49***	.83***
M_{Av} (*s*)		.38***		.44***				.19**	.37***
HS (*s*)			.47***		.53***			.54***	.79***
FF (*s*)				.43**		.68***		-.24***	.70***

Note. P=performance; M=mastery; HS=hope for success; FF=fear of failure; Ap=approach; Av=avoidance.
*p <.05. **p< .01. ***p< .001.

Table 5
Correlations and Significance of the Latent Factors in the Models with Different Grades

	Math grades		Physics grades		German grades	
	r	r_c	*r*	r_c	*r*	r_c
Approach	-.14	-.07	-.21**	-.14	-.04	.03
Avoidance	.06	.09	-.09	-.06	-.07*	-.04
Hope success	-.09	-.09	-.04	-.04	-.14	-.14
Fear of failure	.21***	.20***	.22*	.20*	.03	.01
Mastery	-.11***	-.09***	-.08*	-.06***	-.04	-.01
Performance	-.06	-.05	-.01	.01	-.06	-.05
Global	-.08	-.07	-.06	-.06	.12	.13
Domain	-.45**	-.47***	-.05*	-.06**	.12***	.10***

Note.*N*=325, *r* = zero-order correlations, r_c=semi-partialcorrelations (grades adjusted for conscientiousness).
*$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

Figure Captions

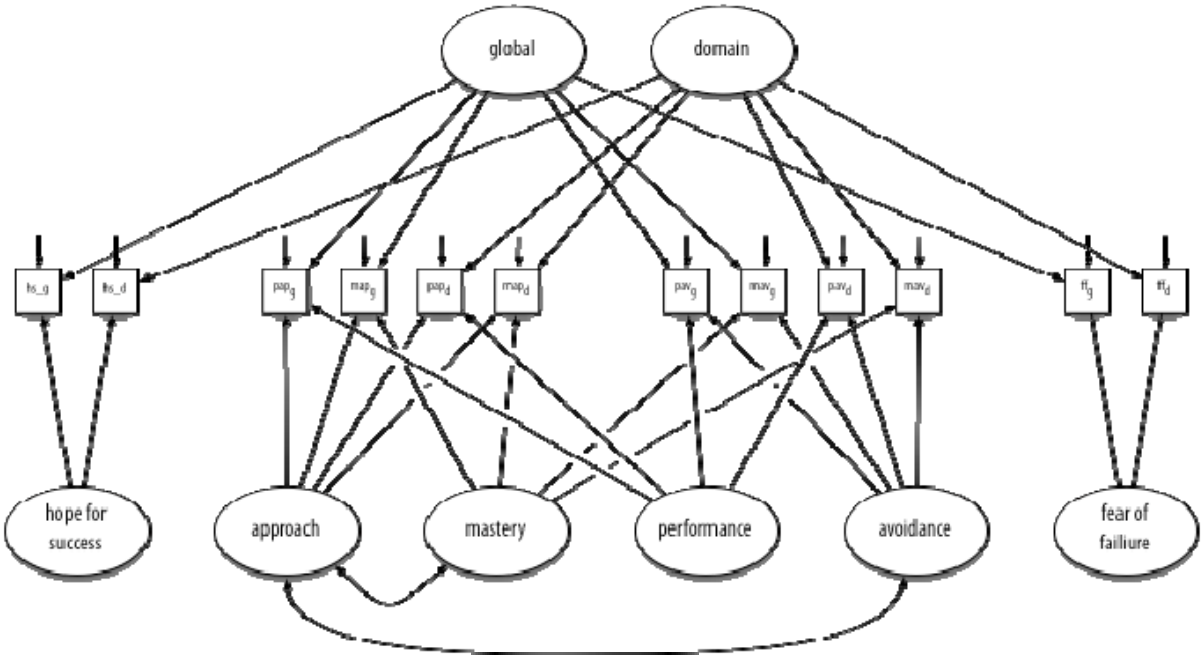

Figure 1. Basic MTMM model.

Is There Really a Single Factor of Personality? A Multirater Approach to the Apex of Personality

Erik Danay and Matthias Ziegler

Humboldt Universität zu Berlin

Author Note

Correspondence should be addressed to Erik Danay at the Psychological Institute, Unter den Linden 6, 10099 Berlin, Germany. E-mail: erik.danay@hu-berlin.de, Telephone: +49 30 2093 9447, Fax: +49 30 2093 9361.

Abstract

A GFP is controversially debated on regarding the hierarchical structure of personality. Support for the existence of such a factor comes from a methodological and a theoretical argumentation. The first calls on the intercorrelations repeatedly found between personality domains. The latter uses evolutionary *K*-theory to advocate the existence of one underlying factor of fitness-enhancing traits. The current study examined both arguments by using a new multirater approach able to disentangle different rater biases and to correct for nested data. Results did not support the idea of a GFP. Furthermore, the bias-adjusted variance of the GFP has a negative relationship with intelligence, making a fitness-enhancing function highly implausible. Findings are discussed including an interpretation of the GFP as successful impression management.

Keywords: GFP, structure of personality, *r*/*K*-theory, plasticity, stability, multilevel, peer ratings, self-ratings, multirater, intelligence, impression management

Is There Really a Single Factor of Personality? A Multirater Approach to the Apex (GFP) of Personality

After an initially turbulent discussion in the 1990s, most personality researchers seemed to agree about the Big 5 framework (Goldberg, 1990). This model, which was derived from a psycho-lexical approach starting in the 1920s (Allport & Odbert, 1936; Klages, 1926), divides personality into five factors. These factors were thought to represent the highest level of personality (McCrae & Costa, 1996). Nevertheless, one of the main criticisms of the Big Five was that there were always substantial correlations between the five factors, despite the claim of orthogonality(Digman, 1997). This was seen as a hint that the Big Five might not represent the apex of personality, but rather some intermediary level. Hence, the General Factor of Personality (GFP) was devised and is now a hot topic among personality researchers. Supporters see in it a concept similar to the *g* factor of intelligence and try to link it to evolution and fitness, whereas opponents are rather skeptical about such a concept and point to methodological problems and possible artifacts at the core of this superfactor. The current study set out to inspect the GFP and determine the possibility of its existence and the viability of its alleged role in human behavior using a new methodological approach based on a multi-informant sample.

The General Factor of Personality (GFP)

The idea of a general factor is nothing new in psychology, the most prominent such factor being undoubtedly Spearman's *g* of intelligence (1904). With such a concept comes the perspective of hierarchy(McGrew, 2009). This means that there are subconstructs of intelligence, most often named as facets on lower levels, domains on higher levels, higher-order factors at an even higher level of abstraction, and at the top, one general factor. By contrast, in personality psychology, the concept of one factor superseding all other domains of personality appears to be quite new. In 1915, Webb found one factor, orthogonal to intelligence's *g*, in observer rater data. This factor, which Webb named *w* for will or volition, was derived by using ratings of the "character side of mental activity" (p. 58). This has sometimes been taken as the first emergence of one general factor of personality, even though Webb's data, when reanalyzed, were more in line with the Big Five view (Deary, 1996) or one domain of the Big Five (i.e., Conscientiousness; (Digman & Inouye, 1986).

Recently, there has been an invigorating push to the notion of a GFP also known as the Big One(Musek, 2007), at the apex of personality, with numerous studies covering this area (e.g., Musek, 2007; Rushton & Irwing, 2008, 2009c; Rushton, Irwing, & Booth, 2010; van der Linden, te Nijenhuis, & Bakker, 2010; Veselka et al., 2009; Zawadzki & Strelau, 2010). The concept of such a factor has been backed up by two lines of argumentation: (a) a methodological line and (b) a theoretical line.

The methodological line of argumentation takes its onset from the substantive intercorrelations of personality domains found consistently in various measures (Digman, 1997). This is especially true for the Big Five that were derived by factor analysis in the first place. Correlated factors, if not due to measurement error or other artifacts, point to another factor one level higher in the hierarchy. Consequently, to support this argumentation, researches began to reanalyze the data from personality inventories in search of the GFP based on the intercorrelations found on the intermediary domain level (e.g., Musek, 2007; Rushton & Irwing, 2008, 2009a, 2009b, 2009c, 2009d; Veselka et al., 2009). In order to do so, they mostly used data from manuals of published personality inventories. Whereas this is

a good way to get large sets of data containing established measures of personality, such an approach nonetheless has one problem: The covariance matrices that the analyses were based on were quite different one from another. Thus, even though a general factor emerged in all of these analyses, it is highly implausible that this factor captured the same variance each time because the "material" was different in each analysis. This speck of doubt was further strengthened by the different loading patterns found in different analyses when using the same traits. Additionally, under the GFP, the structure also varied from two intermediary factors to three or even four factors (e.g., Rushton & Irwing, 2009c). Furthermore, the variance accounted for by the GFP in these analyses ranged roughly between 30 – 60%. This not-so-trivial difference is also a rather interesting finding, which warrants further investigation of the substance of the GFP. Skepticism on the soundness of the concept of a GFP was recently also expressed by Zawadzki and Strelau(2010) who analyzed 32 traits out of six personality inventories in search of a GFP and concluded that there was no real—personality and not temperament-driven—GFP in their data.

For a theoretical line of argumentation supporting the idea of a GFP, Rushton recurred on the evolutionary *r/K*-selection theory (MacArthur & Wilson, 1967), which he had adopted from biology to explain individual differences (Rushton, 1985). *r* refers to the maximal rate of population increase (i.e., proficiency in producing offspring). *K* refers to the capacity of the environment to sustain a certain size of population of a species. In nonhostile environments with unlimited resources to sustain life, species should adopt an *r* strategy for reproduction. This means time should be expended for the purpose of reproduction, but not for looking after the offspring because there is no need for that. In structurally opposite environments (i.e., with innumerable enemies and few resources to be competed for), species should adopt a *K* strategy for reproduction. This means that they should limit themselves to very few offspring and take maximum care of them. Such a strategy demands the most efficient use of the scarce resources available. Comprehensibly, such an environment and such a strategy puts higher demands on attention, intellectual processing (e.g., of danger), allocation of time for crucial tasks, and so forth. Under such a premise, Rushton (1985) formed his Differential *K* Theory and argued that the appearance of people with higher intelligence would rather be favored by the necessity of a *K* environment than that of people with lower intelligence (Rushton, 2004). By doing so, he expanded the idea—which, in the beginning was apparently limited to interspecies differentiation—to intraspecies differentiation. The differences on this *r/K* continuum, as already labeled so by Pianka(1970), could originate from one underlying "basic dimension – K" (Rushton, 1985, p. 445). Apart from intelligence, the personality traits associated with such a strategy are introversion, behavioral restraint, agreeableness, rule-following, altruism, and in general, a lower overall activity level, and of course, sex drive. All these traits combined would form the GFP, a general factor of fitness and, thus, goodness. Recently, Rushton, Bons, and Hur(2008) linked *r/K* theory to heritability, and thus personality, and ultimately the GFP to "efficient" and "inefficient" people (p. 1183). Of course, one could argue that personality and intelligence should not overlap. However, Ackerman and Heggestad (1997) found small correlations on domain level between personality and intelligence. If the GFP would bundle all those variances responsible for these small correlations, then this would be in line with Rushton's argumentation.

This line of argumentation is not without its critics, coming under fire from both psychologists (e.g., Wicherts, Borsboom, & Dolan, 2010) and nonpsychologists(e.g., Graves, 2002). Apart from the debate on the possibility of one general factor, the assumption of heritability has also been questioned by casting doubt on the concept of genetically driven formation of populations (Long & Kittles, 2003). To shed more light on this question, it is

indispensable to determine what exactly is inside this GFP, that is: What is the substance of the GFP?

The Search for Substance in the GFP

After Digman (1997) had brought the idea of higher-order factors of personality into the community's horizon, researchers soon tried to validate the claim of substance and to simultaneously define the content of this substance and connect it to specific behavior. Digman had found two factors, one being constituted by Agreeableness, Conscientiousness, and Emotional Stability, and one being constituted by Extraversion and Openness. Subsequently, the first factor was dubbed by DeYoung, Peterson, and Higgins (2002)*Stability*, the second factor *Plasticity*. Whereas Stability helps people to "get along" with others or to function within straightforward situations (Mount, Barrick, Scullen, & Rounds, 2005, p. 469), Plasticity urges people to broaden their horizons, to evolve, grow, and get ahead in new and complex situations. These definitions, even though providing convincing content-related explanations, did not disperse the doubt about the substantial core of these higher-order factors. That is, "are the correlations among the Big Five real" (DeYoung, 2006, p. 1138) or are they due to rater bias? Since the GFP explains the common variance of Stability and Plasticity, the same questions can be raised. With mono-method studies, that is, mono-rater studies, as forwarded byMusek (2007), for example, this problem has not been overcome (Anusic, Schimmack, Pinkus, & Lockwood, 2009; DeYoung, 2010). Hence, Biesanz, and West (2004) and DeYoung (2006) used a multirater approach to tackle this problem.

The Multirater Approach

Biesanz and West(2004) used self-, peer, and parent ratings of the Big Five traits to test different multitrait-multimethod models. They intentionally used an adjective-based scale, the trait-descriptive adjectives (TDA) by Goldberg (1992), to rule out that a specific Big Five instrument is responsible for making the correlations emerge. In their multitrait-multi-informant models, the correlations between the Big Five did not exist. However, when using amultitrait-multioccasion model (i.e., just one rater), the correlations were there. Therefore, they concluded that the correlations were caused by rater bias. DeYoung (2006) argued against this conclusion by stressing the fact that interrater agreement in Biesanz and West's data was very low (correlation mean = .30), possibly due to the fact of using a single adjective-based scale. This would have made it quite difficult for the Big Five to correlate in the first place. Consequently, DeYoung used one questionnaire with longer items (BFI; John, Donahue, & Kentle, 1991) and one with short items (Mini-Markers; Saucier, 1994) to put the multitrait-multirater models to another test. In his data, interrater agreement was a bit higher (mean correlation = .41 for the BFI, mean = .36 for the Mini-Markers) and comparable with the average inter-rater agreement for family and friends according to a recent meta-analysis by Connelly and Ones (2010). In the models with four different informants, the correlations among the Big Five appeared, and furthermore, a model with Stability and Plasticity at the top of the Big Five achieved an adequate fit. Stability and Plasticity, however, were not correlated, suggesting that they are the apex of personality structure.

These conflicting results are indeed puzzling at first sight. But one explanation of these incompatible findings would be that neither study took into account the nestedness of the data. That is, peer raters are not independent, but are rather nested in one specific target (i.e., the person they have to rate). Thus, defining one "rater variable" (i.e., method variable) that should capture the variance due to pertaining to one specific class of raters (i.e., self, peer,

parent, etc.) does not take into account the hierarchical structure of the data at hand. Ignoring this nested data structure can lead to biased and unreliable covariance estimates. It is therefore recommended to use multilevel models to test hypotheses based on nested data.

The Correlated-Trait Correlated-Method-1 (CTCM-1) Approach

Eid and colleagues (2008) published a paper in which they presented different structural equation models suitable for multitrait-multimethod data. As the most prominent method, they targeted raters. Hence, when referring to multimethod models, different classes of raters are intended. They identified the correlated trait-correlated method model (CTCM) and the correlated trait-correlated uniqueness model (CTCU) as the two most often-applied models for multimethod data. Both such models were used by Biesanz and West (2004), for example. However, for nested data, a CTCM model would only be adequate if "(1) the number of … (raters) did not differ between targets, (2) the loading patterns on the trait and method factors did not differ between the different rater groups, and (3) the method factors were uncorrelated between raters. Hence, imposing these restrictions on the CT-CM model would make the model suitable for the case of interchangeable raters" (Eid et al., 2008, p. 251). Apart from number one, these prerequisites can hardly be met by different rater groups. For example, it is quite plausible that peers base their ratings of another person's personality traits on more visible behavior or verbally expressed sensations than does the person her/himself. Therefore, it is improbable to get the same loading patterns from different rater groups.

Interchangeable versus StructurallyDifferent Raters

Eid and colleagues pointed out that before doing a multimethod analysis it is important to assess whether each method (i.e., rater) is interchangeable or structurally different. Interchangeable raters, for example, do not differ in their perspective on the ratee. On the other hand, structurally different raters would engage with the ratee in and know the ratee from noncomparable situations, and therefore, would have different perspectives. This distinction is important when modeling a rater effect. For peer ratings (i.e., from *peers* and not parents or family-members), it can be assumed that raters are interchangeable. A group of friends knows a person in similar situations. Self-ratings, however, are structurally different from peer ratings because they use a completely different perspective and are, ideally, based on a complete sample of situations.

The CTCM-1 Model for Structurally Different and Interchangeable Raters

If data are composed just of one self- and one peer rating, raters are structurally different. In the case of two peer ratings and one self-rating, we have a combination of structurally different and interchangeable raters. The two peers are interchangeable, the self, however, is structurally different. Furthermore, the peer ratings are not independent, but nested in the self-ratings. In multilevel terms, peer ratings are on level 1, whereas self-ratings, the *target* of the peer ratings, are on level 2. This is illustrated in Figure 1. When using data with structurally different raters, Eid(2000) showed that it is advantageous if one of these methods is defined as the reference method. Thus, he named this model the *correlated trait-correlated method minus 1* because one method is the standard of comparison for which no method factor will be specified. In data in which one method is self-report, it is sensible to define this self-report as the reference method. Hence, method factors for the peer reports reflect the under- or overestimation of these peer ratings with respect to the self-ratings. Furthermore, a

trait factor made out of the self-ratings and the method-adjusted peer ratings contains the *true score* of the trait without any bias from either self or peer.

Applied on the Big Five, these bias-free trait variables, then, should offer the right basis for determining whether there is a GFP after controlling for different rater biases. Consequently, the aim of the current study was twofold:

1. We wanted to establish a hierarchical model of personality with the GFP at the apex with multirater data using an SEM approach that takes into account the nestedness of the data. By doing so, we would make sure that distinct rater biases were controlled for.
2. We wanted to determine whether the GFP has this positive, fitness-enhancing power as it should according to Differential *K* theory. In order to do so, we correlated this GFP with an established measure of general intelligence.

Method

Sample and Procedure

Fourhundred six students of social sciences were recruited as participants (82% female). Their mean age was 22.67 (*SD* = 5.84). As a requirement for participation, each participant had to name two peers who would be able to act as raters. As an incentive to hand in the ratings by the two peers, participants were offered detailed feedback on their results as soon as they had delivered the peer ratings. Because it was necessary that peer raters had a similar history of knowing the ratee, as a prerequisite, raters had to be friends with the ratee for at least 2 years.

The tests were administered in a lab with a maximum of eight participants at one time. After completing the test, participants were handed two envelopes containing the questionnaire for the peer ratings along with instructions on how to compile the questionnaire. For seven participants we obtained only one peer rating instead of two (1.7%).

Test Materials

Personality self-ratings. The German version of the NEO-PI-R (Ostendorf & Angleitner, 2004) was used. The NEO-PI-R consists of 240 items, with 8 items for each of the six facets of the factors of the Five Factor Model (i.e., Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness). Items require participants to rate themselves in typical behaviors or reactions on a 5-point Likert scale ranging from 0 (*strongly disagree*) to 4 (*strongly agree*). Descriptive statistics as well as reliabilities for the scale can be found in Table 1.

Personality other ratings. For peer ratings, again, the NEO-PI-R was used, but in a shortened version in order to increase compliance. For each facet, only one item was used, bringing the total sum of items down to 30. Items consisted of adjectives taken from the description for each facet in the manual. The rating scale was the same 5-point Likert scale as for the self-ratings. The inter-rater agreement correlation across all raters was r=.44, split for each target, the average r=.45 (*SD*=.23). As a better measure of inter-rater agreement, an ICC1 was computed separately for each target. The average of this ICC, which takes absolute agreement into account, was .41 (*SD* = .24). These values for ratings by friends were in the range of values presented in the meta-analysis by Ones and Connelly (2010).

Intelligence. Cognitive ability was assessed with the basic module of the Intelligence Structure Test 2000 R (Amthauer, Brocke, Liepmann, & Beauducel, 2001). The IST-2000-R is a well-established measure of intelligence providing scores for verbal (Cronbach's α = .79, theoretical score range 0 – 60, *M* = 39.29, *SD*= 6.29), numerical (Cronbach's α = .92, theoretical score range 0 – 60, *M* = 41.90, *SD*= 9.57), and figural intelligence (Cronbach's α = .84, theoretical score range 0 – 60, *M* = 35.61, *SD*= 7.96). The global score can be interpreted as a general reasoning score (Cronbach's α = .93, theoretical score range 0 – 180, *M* = 116.81, *SD*= 18.82) used in the analyses here.

Statistical Analyses

The CTCM-1 model as depicted in Figure 1 was tested in MPlus 5.2 (Muthén&Muthén, 1998-2007) using a robust ML estimator. Cutoffs used to assess the model fit were based upon the suggestions by Beauducel and Wittmann (2005) and Hu and Bentler (1999). Consequently, we looked at the χ^2 value, the SRMR (which should be below .11), the RMSEA (which should be less than or equal to .06), and the CFI (which should be approximately .95).

In Model 1, trait variables Neuroticism, Agreeableness, and Conscientiousness should form the Stability factor whereas Openness to Experience and Extraversion should form the Plasticity factor as described by DeYoung and colleagues (2002). Situated above these factors is the GFP. This model was used to test the existence of the GFP. Model 2 served to test whether the GFP is related to intelligence. In order to do so, we correlated the GFP with measures of intelligence. In particular, we used the reasoning score in the IST-2000-R. Reasoning is thought to be the best indicator of fluid intelligence (Carroll, 1993). Additionally, we used the three basic facets of intelligence (i.e., verbal, numerical, and figural intelligence) to get an even better-differentiated picture. Therefore, we regressed the GFP onto the three facets. All three facets were correlated in order to establish the unique influence of each intelligence facet.

Results

Model 1 had the following global fit: $\chi^2(39) = 194.5$, $p < .001$. However, considering the suggestions by Beauducel and Wittmann (2005), fit indices were in the expected range for questionnaire data: RMSEA = .07, CFI = .851, $SRMR_{within}$ = .018, $SRMR_{between}$ = .123.

Loadings and significance levels can be found in Figure 1. In order to get a convergent model, we had to fix the error variance of the latent Openness factor to 1. Therefore, no estimates for the standardized loadings were computed. The loading from the trait variable Extraversion (λ = .42) on the higher-order factor Plasticity was significant with $p < .001$. Loadings from the trait variables Neuroticism (λ = -.37), Agreeableness (λ = .67), and Conscientiousness (λ = .24) on the higher-order factor Stability were also all significant with $p < .01$. Loadings from Plasticity (λ = .45) and Stability (λ = .45) on the GFP were also both significant with $p < .001$. However, the GFP on top of these two factors did not have significant variance, which suggests that it does not exist. Construct reliability (Hancock & Mueller, 2001) was also quite low for the GFP with *H* = .34. This coefficient *H* is also referred to as construct replicability (i.e., the stability of a construct) as reflected in the data on the chosen indictors.

Even though the GFP did not appear in Model 1, this could also have been a problem of power. Hence, Model 2 was used to test one of the GFP theorists' fundamental claims, that is, the fit-enhancing valence of the GFP and thus, its relationship with intelligence. Therefore, the GFP was first correlated with reasoning. Model 2 had the following fit: $\chi^2(48) = 218.84$, $p < .001$, with fit indices showing acceptable values: RMSEA = .066, CFI = .839, $SRMR_{within}$ = .016, $SRMR_{between}$ = .121. The GFP correlated with the reasoning score $r = -.10$, $p = .44$. Thus, there was no significant connection between the GFP and fluid intelligence. Furthermore, the direction of the correlation would have been against theory (i.e., negative).

In an additional step, we regressed the GFP on the three intelligence facets verbal, numerical, and figural intelligence. In doing so, the unique influence of each intelligence facet apart from their shared variance (reasoning) would emerge. The model fit was as follows: $\chi^2(67) = 248.74$, $p < .001$; RMSEA = .06, CFI = .852, $SRMR_{within}$ = .014, $SRMR_{between}$ = .111. The regression weights from the facets of intelligence in predicting the GFP were not significant for numerical and figural intelligence, whereas the regression weight for verbal intelligence ($\beta = .30$) did reach significance ($p = .028$).The variance explained in the GFP by the three intelligence facets was $R^2 = .14$. Because the nonsignificant weight for numerical intelligence was negative, we checked the zero-order correlation between the GFP and numerical intelligence in order to rule out a suppression effect. The zero-order correlation was also negative and not significant as well. A suppression effect could therefore be ruled out.

Discussion

The current study tried to establish whether there is a GFP at the apex of personality through multirater data modeled by a CTCM-1 approach. Results did not support such an idea. Furthermore, the GFP was claimed to be associated with positive, fit-enhancing traits based on assumptions made in differential *K* Theory (Rushton, 1985). Accordingly, it should covary with one of the most fit-enhancing traits in mankind: intelligence. Results showed that the GFP had a negative, though statistically nonsignificant correlation with reasoning. This is against the most fundamental theoretical argument for the existence of the GFP (i.e., that fit-enhancing traits co-evolved and form a *K* factor). Additionally, when regressed onto verbal, numerical, and figural intelligence, only verbal intelligence had a significant and positive impact on the GFP. Even more, numerical intelligence had a reversed (i.e., negative) impact on the GFP. The overall explained variance was moderate. This means that the GFP, if it had existed, would be positively connected with verbal intelligence, but negatively related to numerical and figural intelligence. In other words, the higher someone's GFP, the higher her/his verbal IQ, but the lower her/his numerical IQ. This makes the idea of the GFP as a *generally* fit-enhancing factor implausible.

Given these results, we have to turn to the question of what made something like the GFP appear in so many sets of data. Two possible explanations were mentioned above: substance or bias.

The GFP as Substance

Those in support of the GFP fall back on Differential *K* theory to explain the variance captured by this one factor. According to this theory, throughout evolution, personality traits have covaried with, among other traits, intelligence, altruism, introversion, and reproductive strategies. The GFP should capture the common variance in these desirable traits and form one underlying trait: the *K* factor. Based on our data, however, this assumption had to be

rejected. First, in our data, people at the positive pole of the GFP were less introverted and exhibited less reasoning. Even more so, those with higher verbal intelligence scored higher on the GFP, whereas those with higher numerical intelligence scored lower. This is quite puzzling considering that intelligence facets are usually positively correlated around $r = .60$ (McGrew, 2009). The inconsistency of the influence of the facets of intelligence points to the fact that whatever is captured in the GFP is not uniformly evolutionarily fit enhancing. For example, it is hardly plausible that talking would be evolutionarily advantageous while simultaneously, calculating would not be, especially given the fact that language was one of the latest features to evolve in humans. The same reasoning is equally valid for spatial coordination and mapping (i.e., figural intelligence), which again, was found to have only a negligible influence on the GFP. However, in evolutionary environments, such a skill was vitally important (e.g., when hunting). Taken together, these ideas lead to the conclusion that the substance in the GFP—if there was one—could not be explained by the *K* factor and its proposed evolutionarily advantageous influence.

The GFP as Bias

The other possible explanation for the GFP is variance due to artifact or bias. Since the emergence of higher-order factors above the Big Five (Digman, 1997), some researchers have held measurement error, and in particular rater bias, to be responsible for the materialization of these factors. The argument that advocates of the GFP have used to counter this view is that when controlling for bias, there is still substantive variance left to account for a GFP (Rushton et al., 2010). However, when using an experimental design, this claim has already been disproven (Ziegler & Bühner, 2009). Moreover, none of the studies cited in support of the argument that the GFP is more than bias have used a multilevel-multirater approach to adequately handle the data and possible biases. DeYoung (2006) has already suggested that if one could reliably show that the correlations are due only to bias, "all discussion of higher-order personality factors above the Big Five would be pointless, except inasmuch as one is interested in systematic biases in personality perception" (p. 1139). In the CTCM-1 approach used in the present study, we controlled for all distinct rater biases. Thus, all variance ending up in the latent trait variables had to be shared by all individual raters. This common variance could be either true score variance or some other construct-independent variance that was, nonetheless, shared by all raters. Construct-independent systematic variance is normally referred to as bias (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). Because of the necessity of being shared by all raters, the bias left in the CTCM-1 model is in a special category.

There are two possible explanations for such a bias. First, this could be bias due to socially desirable responding (SDR). Such a bias was defined as the result of a person-by-situation interaction (Heggestad, George, & Reeve, 2006; Ziegler, Toomela, & Bühner, 2009). That is, SDR is activated in specific situations and it does not happen indiscriminately for all traits (Birkeland, Manson, Kisamore, Brannick, & Smith, 2006) in any given situation. Looking at how self-ratings and peer ratings were obtained, such an explanation for common bias is highly implausible. Self-ratings were obtained in a completely different situation (i.e., a lab at the university) than the peer ratings (at home). For these situations to be similar is not very plausible. Even the peer ratings themselves were most likely not given in similar situations.

Ruling out general SDR, the bias supposedly captured by the GFP must be a bias that is inherent in all ratings. The question is, how can the same bias manifest itself in the different ratings investigated here? Normally this is reached when the raters share the same view. For people to have the same view on things, these things have to be either (a) easily observable

(John & Robins, 1993; Vazire & Mehl, 2008), or (b) frequently referred to or talked about. The first would speak against a bias. The second is most easily accomplished when traits are disclosed through self-description. Such disclosure, of course, may be moderated by the importance a certain trait has for oneself or socially (Koestner, Bemieri, & Zuckerman, 1994). In any case, the person talking about specific behavior he or she actually never exhibits or showing behavior out of the person's nature and doing both in a convincing way has at least one distinct skill: to successfully present her/himself in the way she or he wants to appear to others. Hence, such a person is highly successful at impression management. Both paths to convince others of one's self can be actively influenced by the ratee by again and again showing a specific behavior or talking about it. Both can be done consciously or unconsciously. However, when looking closer at both ways of convincing, for the first way (i.e., showing a specific behavior), it cannot be ruled out that this behavior also represents the true trait standing of the person because that is how personality is defined: as observable behavior in specific situations. This makes the other way of setting the ground for a common rater bias more likely: oral description. For oral accounts to be convincing, it seems plausible that verbal intelligence provides an advantage. This is backed up by our data. Moreover, to be able to tell a story for others to readily listen to, one has to be pleasant company. When Rushton and colleagues (2008) explained why the underlying dimension of the GFP would evolve over time through more and fitter offspring, they called to attention that "people prefer as mates, fellow workers, and leaders, those who are *agreeable*, *cooperative* and *emotionally stable*" [emphasis added](cf., Figueredo & Rushton, 2009; Rushton et al., 2008, p. 1181), thus making such behavior the selection criteria. Therefore, a different explanation for the GFP would be that people with high scores are able to present themselves as agreeable, cooperative, and emotionally stable. In congruence with such reasoning, loadings from Agreeableness were highest in our model, followed by Extraversion and Emotional Stability (Neuroticism). This loading pattern, as was mentioned above, is against the assumptions of Differential *K* theory, which claims that people with higher GFP scores are less extraverted.

Altogether, a reconciliation of these two explanations – bias versus substance – can be reached: If the variance inside the GFP is not a substantial personality trait in the strict sense, it may possibly be a skill best referred to as successful impression management. Such a skill could very well have evolved across human evolution. Its implications, however, are quite different from those originally forwarded by *K* theorists. This would not be something to be associated with a personality trait in the strict sense of the word because it seems to be connected very much with social skills and the ease with which a person engages in social contact. It should also not be a genetically evolved population-based discriminant, which would have to be uniform at its base. Using the argument from above that genetic selection may possibly happen not so much at the population level, but rather at a family or kin level, such a perspective can help us to better understand such a concept because although it is true that social skills are based on personality traits, even more so, they are based on acculturation and learning (for a most recent overview of biology research in this area, see Laland, Odling-Smee, & Myles, 2010). Such learning most naturally happens in families and most often so by watching elders or parents in their daily habits and by copying and adapting to how they do things. It is therefore conceivable that differences in the GFP are far greater between families than between populations. But it is also conceivable that different social displays and interaction styles may give different composites of variance for such a skill. Extraversion may, for example, not be as necessary for a well-functioning social interaction in a culture such as Japan or the likes. Similarly, such a skill may not be irreversibly cast in stone but could possibly be honed by gathering verbal knowledge and interlocutory experience and, as such, evolve over time and maturation. But, as pointed out above, this would not be

imperative because a more introverted person (i.e., one who prefers doing things on his or her own) would not need to be less fit; the path to obtaining fitness, however, would be a different one: Such a person would focus the deployment of resources on a more closed circle of benefitors and hence look for a more directly rewarding, self-centered investment. From an evolutionary perspective, however, both approaches could be successful. This is even more true when taking into account that most concepts in psychology are not bipolar but rather unipolar. Such an understanding seems reasonable for the skill of successful impression management as well. Everyone will to some extent manage the impressions she/he makes on others. The reach of such an impression (e.g., its stability and power), however, may vary.

Limitations

A limiting factor to the study was the use of only one personality questionnaire, even though the questionnaire that was employed ranks among the most popular in research and praxis. Another limiting aspect was the character of the sample, which included only students. This may have made hindered the occurrence of cross-loadings and thus, correlations between the Big 5 domains.

Outlook

The present study did not support the idea of a substantial GFP at the apex of personality as promoted. Even more so, the claim of the GFP to be an underlying dimension of an evolutionarily positive trait as suggested by *K* theory could not be confirmed. Instead, results made it plausible that the GFP, if really existent, is due to bias and most probably to successful impression management. Accordingly, people with greater verbal skills are those at the positive pole of the GFP. This explanation does not rule out the evolutionary importance of the GFP but sheds a completely different light on its theoretical underpinnings. This aspect should be taken as a starting point for future research in trying to link impression management to the GFP. This would also give further insight into the substance of the intermediary factors Stability and Plasticity, which seem to have substance even after controlling for rater bias.

References

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. Psychological Bulletin*, 121*, 219-245.

Allport, G. W., & Odbert, H. S. (1936). Trait names: A psycho-lexical study. Psychological Monographs*, 47 (1 , Whole No. 2 1 1).*

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). I-S-T 2000 R (Intelligenz-Struktur-Test 2000 R) [Intelligence-Structure-Test 2000 R]. Göttingen: Hogrefe.

Anusic, I., Schimmack, U., Pinkus, R. T., & Lockwood, P. (2009). The Nature and Structure of Correlations Among Big Five Ratings: The Halo-Alpha-Beta Model. Journal of Personality and Social Psychology*, 97*(6), 1142-1156.

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. Structural Equation Modeling*, 12*(1), 41-75.

Biesanz, J. C., & West, S. G. (2004). Towards Understanding Assessments of the Big Five: Multitrait-Multimethod Analyses of Convergent and Discriminant Validity Across Measurement Occasion and Type of Observer. Journal of Personality*, 72*(4), 845-876.

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.

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies: New York: Cambridge University Press.

Connelly, B. S., & Ones, D. S. (2010). An Other Perspective on Personality: Meta-Analytic Integration of Observers' Accuracy and Predictive Validity. Psychological Bulletin*, 136*(6), 1092-1122.

Deary, I. J. (1996). A (latent) big five personality model in 1915? A reanalysis of Webb's data. Journal of Personality and Social Psychology*, 71*(5), 992-1005.

DeYoung, C. G. (2006). Higher-order factors of the Big Five in a multi-informant sample. Journal of Personality and Social Psychology*, 91*(6), 1138-1151.

DeYoung, C. G. (2010). Toward a Theory of the Big Five. Psychological Inquiry*, 21*(1), 26-33.

DeYoung, C. G., Peterson, J. B., & Higgins, D. M. (2002). Higher-order factors of the Big Five predict conformity: Are there neuroses of health? Personality and Individual Differences*. 33*(4), 533-552.

Digman, J. M. (1997). Higher-order factors of the Big Five. Journal of Personality and Social Psychology*, 73*(6), 1246-1256.

Digman, J. M., & Inouye, J. (1986). Further Specification of the 5 Robust Factors of Personality. Journal of Personality and Social Psychology, *50*(1), 116-123.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. Psychometrika, *65*(2), 241-261.

Eid, M., Nussbeck, F., Geiser, C., Cole, D., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods. Psychological Methods, *13*(3), 230-253.

Figueredo, A. J., & Rushton, J. P. (2009). Evidence for Shared Genetic Dominance Between the General Factor of Personality, Mental and Physical Health, and Life History Traits. *Twin* Research and Human Genetics, *12*(6), 555-563.

Goldberg, L. (1990). An alternative "description of personality": The Big-Five factor structure. Journal of Personality and Social Psychology, *59*(6), 1216-1229.

Goldberg, L. (1992). The development of marker variables for the Big-Five factor structure. Psychological Assessment, *4*, 26-42.

Graves, J. L. (2002). What a tangled web he weaves. Anthropological Theory, *2*(2), 131-154.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. In R. Cudeck, S. d. Toit & D. Sörbom (Eds.), Structural Equation Modeling: Present and Future - Festschrift in honor of Karl Jöreskog (pp. 195-216): Lincolnwood, IL: Scientific Software International, Inc.

Heggestad, E. D., George, E., & Reeve, C. L. (2006). Transient error in personality scores: Considering honest and faked responses. Personality and Individual Differences, *40*(6), 1201-1211.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, *6*(1), 1-55.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). The Big Five Inventory: Versions 4a and 54, Institute of Personality and Social Research. *University of California, Berkeley, CA*.

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. Journal of Personality. *61*(4), 521-551.

Klages, L. (1926). Die Grundlagen der Charakterkunde [Science of character]. Leipzig: Barth.

Koestner, R., Bemieri, F., & Zuckerman, M. (1994). Self-Peer Agreement as a Function of Two Kinds of Trait Relevance: Personal and Social. Social Behavior and Personality: an international journal, *22*, 17-30.

Laland, K. N., Odling-Smee, J., & Myles, S. (2010). How culture shaped the human genome: bringing genetics and the human sciences together. Nature Reviews Genetics, *11*(2), 137-148.

Long, J. C., & Kittles, R. A. (2003). Human Genetic Diversity and the Nonexistence of Biology Races. Human Biology, *75*(4), 449-471.

MacArthur, R. H., & Wilson, E. O. (1967). The theory of island biogeography. Princeton, NJ: Princeton University Press.

McCrae, R. R., & Costa, P. T., Jr. (1996). Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In J. S. Wiggins (Ed.), The five-factor model of personality: Theoretical perspectives (pp. 51-87). New York: Guilford.

McGrew, K. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. Intelligence, *37*(1), 1-10.

Mount, M. K., Barrick, M. R., Scullen, S. M., & Rounds, J. (2005). Higher-order dimensions of the big five personality traits and the big six vocational interest types. Personnel Psychology, *58*, 447-478.

Musek, J. (2007). A general factor of personality: Evidence for the Big One in the five-factor model. Journal of Research in Personality, *41*(6), 1213-1233.

Muthén, L. K., & Muthén, B. (1998-2007). MPlus User's Guide. 5th edition. Los Angeles, CA: Muthén & Muthén.

Ostendorf, F., & Angleitner, A. (2004). NEO-PI-R. NEO Persönlichkeitsinventar nach Costa und McCrae. Revidierte Fassung. [NEO-PI-R. NEO Personality Inventory]. Göttingen: Hogrefe.

Pianka, E. R. (1970). R-Selection and K-Selection. American Naturalist, *104*(940), 592-597.

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. Journal of Applied Psychology, *88*(5), 879-903.

Rushton, J. P. (1985). Differential K Theory: The sociobiology of individual and group differences. Personality and Individual Differences, *6*(4), 441-452.

Rushton, J. P. (2004). Placing intelligence into an evolutionary framework or how fits into the r-K matrix of life-history traits including longevity. Intelligence, *32*(4), 321-328.

Rushton, J. P., Bons, T., & Hur, Y. (2008). The genetics and evolution of the general factor of personality. Journal of Research in Personality, *42*(5), 1173-1185.

Rushton, J. P., & Irwing, P. (2008). A General Factor of Personality (GFP) from two meta-analyses of the Big Five: Digman (1997) and Mount, Barrick, Scullen, and Rounds (2005). Personality and Individual Differences, *45*(7), 679-683.

Rushton, J. P., & Irwing, P. (2009a). A General Factor of Personality (GFP) from the Multidimensional Personality Questionnaire. Personality and Individual Differences, *47*(6), 571-576.

Rushton, J. P., & Irwing, P. (2009b). A General Factor of Personality in 16 sets of the Big Five, the Guilford–Zimmerman Temperament Survey, the California Psychological Inventory, and the Temperament and Character Inventory. Personality and Individual Differences, *47*(6), 558-564.

Rushton, J. P., & Irwing, P. (2009c). A general factor of personality in the Comrey Personality Scales, the Minnesota Multiphasic Personality Inventory-2, and the Multicultural Personality Questionnaire. Personality and Individual Differences, *46*(4), 437-442.

Rushton, J. P., & Irwing, P. (2009d). A General Factor of Personality in the Millon Clinical Multiaxial Inventory-III, the Dimensional Assessment of Personality Pathology, and the Personality Assessment Inventory. Journal of Research in Personality, *43*(6), 1091-1095.

Rushton, J. P., Irwing, P., & Booth, T. (2010). A General Factor of Personality (GFP) in the Personality Disorders: Three Studies of the Dimensional Assessment of Personality Pathology-Basic Questionnaire (DAPP-BQ). Twin Research and Human Genetics, *13*(4), 301-311.

Saucier, G. (1994). Mini-Markers - a Brief Version of Goldberg Unipolar Big-5 Markers. Journal of Personality Assessment, *63*(3), 506-516.

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The* American Journal of Psychology, *15*(2), 201-292.

van der Linden, D., te Nijenhuis, J., & Bakker, A. B. (2010). The General Factor of Personality: A meta-analysis of Big Five intercorrelations and a criterion-related validity study. Journal of Research in Personality, *44*(3), 315-327.

Vazire, S., & Mehl, M. R. (2008). Knowing Me, Knowing You: The Accuracy and Unique Predictive Validity of Self-Ratings and Other-Ratings of Daily Behavior. Journal of Personality and Social Psychology, *95*(5), 1202-1216.

Veselka, L., Schermer, J. A., Petrides, K. V., Cherkas, L. F., Spector, T. D., & Vernon, P. A. (2009). A General Factor of Personality: Evidence from the HEXACO Model and a Measure of Trait Emotional Intelligence. Twin Research and Human Genetics, *12*(5), 420-424.

Webb, E. (1915). Character and intelligence: An attempt at an exact study of character. British Journal of Psychology Monographs, *1*(3), 1-99.

Wicherts, J. M., Borsboom, D., & Dolan, C. V. (2010). Evolution, brain size, and the national IQ of peoples around 3000 years B.C. Personality and Individual Differences, *48*(2), 104-106.

Zawadzki, B., & Strelau, J. (2010). Structure of personality: Search for a general factor viewed from a temperament perspective. Personality and Individual Differences, *49*(2), 77-82.

Ziegler, M., & Bühner, M. (2009). Modeling Socially Desirable Responding and Its Effects. Educational and Psychological Measurement, *69*(4), 548.

Ziegler, M., Toomela, A., & Bühner, M. (2009). A Reanalysis of Toomela (2003): Spurious measurement error as cause for common variance between personality factors. Psychology Science Quarterly, *51*, 65-75.

Table 1
Descriptive Statistics for Self- and Peer Ratings

Domain	M_{self}	SD_{self}	α_{self}	M_{peers}	SD_{peers}	α_{peers}	1.	2.	3.	4.	5.
1. Neuroticism	1.94	.49	.94	2.25	.67	.70	-	.25**	.11**	.27**	.13**
2. Extraversion	2.40	.42	.91	2.89	.59	.67	-.38**	-	.42**	.24**	.10**
3. Openness	2.68	.35	.90	3.10	.59	.74	-.07	.37**	-	.25**	.15**
4. Agreeableness	2.42	.35	.89	2.98	.57	.73	-.10*	.11*	.17**	-	.25**
5. Conscientiousness	2.45	.41	.90	2.91	.65	.83	-.19**	.16**	.03	.12*	-

Note. $n_{\text{self-rating}}$= 406; $n_{\text{peer ratings}}$= 805. M_{self}= Mean value of self-ratings (theoretical range from 0 to 4). M_{peers}= Mean value of average peer ratings (theoretical range from 0 to 4). Values above the diagonal are the correlations for average peer ratings. Values below the diagonal are correlations for self-ratings.
*$p \leq .05$. **$p \leq .01$.

Figure Captions

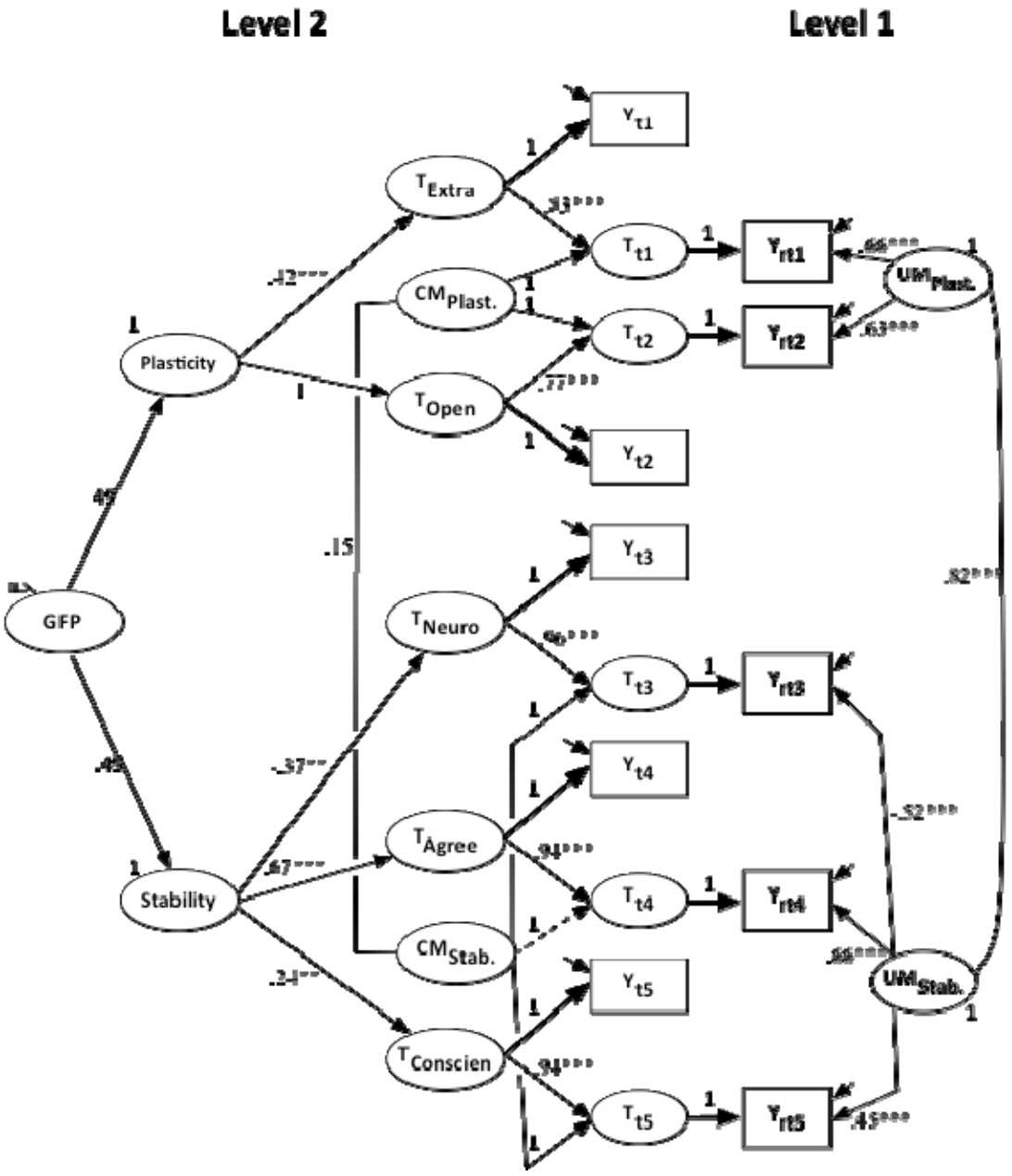

Figure 1: Model 1.

Note. T = trait, CM = Common Method Factor (i.e., by observer raters), UM = Unique Method Factor (i.e., by one specific rater), Y = observed variable, t = trait, r = rater. On Level 1, only one exemplary rater is depicted. Bold paths are fixed to 1 according to the CTCM-1 model.

Danksagung

Die dieser Dissertation zugrundeliegenden Arbeiten entstanden alle am Lehrstuhl für Psychologische Diagnostik von Prof. Dr. Matthias Ziegler am Institut für Psychologie der Humboldt Universität zu Berlin. Prof. Ziegler gebührt auch mein größter Dank. In all den Monaten der Arbeit an der Dissertation waren seine immense fachliche Kompetenz und sein präziser Rat mir eine enorme Unterstützung, sein unermüdlicher Forschergeist mir Vorbild. Wenn es darum geht, das Idealbild eines Forschers und Betreuers zu zeichnen, so kann in meinen Augen vor allem er dazu dienen. Des Weiteren möchte ich Prof. Dr. Anne Frenzel danken, die mich bei der Erstellung von Studie zwei mit Ihrem wertvollen Rat und entscheidenden Hinweisen unterstützt hat. Ebenso möchte ich Prof. Dr. Markus Bühner danken, dass er durch viele gute Hinweise mich an seinem Wissen teilhaben ließ und dass er sich die Zeit und Mühe nahm, mich in meinem Dissertationsprojekt zu unterstützen. Mindestens genauso großer Dank gebührt auch Prof. Dr. Thomas Götz, der ohne zu zögern sich bereit erklärt hat, ebenfalls Teil meines Dissertationsvorhaben zu sein trotz des dadurch so großen Aufwandes an Zeit und Mühen. Ihnen allen gebührt mein Dank.