Psychology Science, Volume 49, 2007 (2), p. 88-114

# Extended criteria and predictors in college admission: Exploring the structure of study success and investigating the validity of domain knowledge

OLGA KUNINA[1], OLIVER WILHELM[1], MAREN FORMAZIN[1],
KATHRIN JONKMANN[2] & ULRICH SCHROEDERS[1]

## Abstract

The utility of aptitude tests and intelligence measures in the prediction of the success in college is one of the empirically best supported results in ability research. However, the structure of the criterion "study success" has not been appropriately investigated so far. Moreover, it remains unclear which aspect of intelligence – fluid intelligence or crystallized intelligence – has the major impact on the prediction. In three studies we have investigated the dimensionality of the criterion achievements as well as the relative contributions of competing ability predictors. In the first study, the dimensionality of college grades was explored in a sample of 629 alumni. A measurement model with two correlated latent factors distinguishing undergraduate college grades on the one hand from graduate college grades on the other hand had the best fit to the data. In the second study, a group of 179 graduate students completed a Psychology knowledge test and provided available college grades in undergraduate studies. A model separating a general latent factor for Psychology knowledge from a nested method factor for college grades, and a second nested factor for "experimental orientation" had the best fit to the data. In the third study the predictive power of domain specific knowledge tests in Mathematics, English, and Biology was investigated. A sample of 387 undergraduate students in this prospective study additionally completed a compilation of fluid intelligence tests. The results of this study indicate as expected that: a) ability measures are incrementally predictive over school grades in predicting exam grades; and b) that knowledge tests from relevant domains were incrementally predictive over fluid intelligence. The results of these studies suggest that criteria for college admission tests deserve and warrant more attention, and that domain specific ability indicators can contribute to the predictive validity of established admission tests.

Key words: College admission, incremental validity, knowledge tests, fluid intelligence, crystallized intelligence.

[1] Olga Kunina, IQB (Institute for Educational Progress), Humboldt-Universität zu Berlin, Jägerstr. 10/11, 10117 Berlin, Germany; email: olga.kunina@iqb.hu-berlin.de

[2] Max Planck Institute for Human Development

## Theoretical background

In many Western countries aptitude tests are accepted and well validated selection criteria for college admission and other high stakes purposes (Camara & Echternacht, 2000). In Germany the use of aptitude tests for college admission is uncommon. Substantial judicial and structural amendments in German student admission politics since August 2004 strengthen the role of universities in the selection process. These changes caused an intensive debate on suitable selection criteria and gave an important impulse for the realization of the studies presented in this article.

## Predicting success in college

The validity of school grades (predominantly high school grade point average (or GPA) - further referred as high school GPA) for the prediction of exam grades at college has been demonstrated in various studies (Schuler, Funke, & Baron-Boldt, 1990; Gold & Souvignier, 2005; Trapmann, Hell, Weigand, & Schuler, in press). Several studies have demonstrated that these results were also valid for Psychology studies in Germany (Schmidt-Atzert, 2005; Steyer, Yousfi, & Würfel, 2005) as well as in the United States (Fenster, Markus, Wiedemann, Brackett, & Fernandez, 2001).

Incremental validity of aptitude tests over school grades was shown for SAT I (Burton & Ramist, 2001; Bridgeman, Jenkins, & Ervin, 2000; Ramist, Lewis, & McCamley-Jenkins, 1993), GRE (Morrison & Morrison, 1995; Kuncel, Hezlett, & Ones, 2001; Burton & Wang, 2005) and ACT (Noble & Sawyer, 2002). The German Medical Entrance Test TMS, which was applied in Germany between 1986 and 1996, also explained additional variance of college grades over school grades (Stumpf & Nauels, 1990; Trost, Klieme, & Nauels, 1997; Trost et al., 1998).

The results from different meta-analytic studies indicate that intelligence measures are strong predictors of study success for different study courses and various criterion information of study success (e.g. Freshman-GPA, cumulative GPA, faculty rankings) (Kuncel, Hezlett, & Ones, 2004; Ones, Visweswaran, & Dilchert, 2005; Kuncel & Hezlett, 2007). In this approach usually general cognitive ability (the “g factor”) is focussed on. However, these validity-generalization studies comprise aptitude tests such as SAT, GRE and MCAT in addition to classical intelligence tests implying that all these tests measure the same ability. Another serious limitation of these meta-analytic studies is lack of distinction between fluid and crystallized intelligence (Cattell, 1987). Cattell (1987) and Horn (1988) described fluid intelligence as a reasoning and problem solving ability which is almost independent from previous learning experiences. In contrast, crystallized intelligence comprises abilities highly influenced by schooling and acculturation. Moderate correlations between crystallized and fluid intelligence varying from $r = .40$ to $r = .50$ (Cattell, 1987; Brody, 1992) indicate that the two aspects are not identical.

The validity of domain-specific knowledge tests such as SAT II and GRE subject tests in the prediction of study success has been less intensively investigated. A simulation of the consequences of applying different predictors in student admission decisions has shown that, given the high correlation of $r = .84$ between SAT I and SAT II, no major differences can be found between both selection models (Bridgeman, Burton & Cline, 2001). Geiser and Stud-

ley (2001) reported that SAT II was the best single predictor in the prediction of Freshman-GPA. In their study, SAT I explained only a small proportion of the variance of college grades after school grades and SAT II scores were taken into account. It was also demonstrated that the corrected validity coefficients for school grades ($r$ = .63), SAT I ($r$ = .60) and SAT II ($r$ = .62) as well as the incremental validity of SAT I and SAT II (over the combination of SAT II and school grades and SAT I and school grades, respectively) were almost identical in the prediction of Freshman-GPA (Ramist, Lewis, & McCamley-Jenkins, 2001). These results are probably consequences of the high collinearity between SAT I and SAT II. The predictive validity coefficients varied broadly for different SAT II subject tests, ranging from $r$ = .17 (for German or Spanish language) to $r$ = .58 (for Mathematics or Chemistry) (Ramist et al., 2001). An extensive review on validity comparison between SAT I and SAT II is provided by Kobrin, Camara and Milewski (2002).

Relating the theoretical conceptualisations of SAT I and SAT II to established intelligence models (Carroll, 1993; Cattell, 1987; Horn, 1988) SAT I predominantly measures individual differences in fluid aspects of intelligence while SAT II primarily reflects individual differences in crystallized aspects of intelligence. These considerations lead to the question why the correlation between SAT I and SAT II ($r$ = .84) is noticeably higher than the moderate correlation between fluid and crystallized intelligence reported above. A detailed inspection of the SAT II tasks reveals that fluid and crystallized aspects are confounded in these items. They require high fluid intelligence to transfer the acquired knowledge to new objectives. This substantial demand of fluid intelligence for successful completion of SAT II items might explain the high relation between SAT I and SAT II.

Similarly to the SAT, the Graduate Record Examinations (GRE) consist of two parts – a reasoning test including verbal, quantitative and analytical writing sections and several subject tests available for 9 majors, including Psychology. In some studies, only small relations between GRE and study success were found. Goldberg and Alliger (1992) report a standardized regression coefficient of $\hat{\beta}$ = .15 for the GRE predicting cumulative GPA in PhD studies. Validity coefficients for the GRE almost double when they are corrected for restriction of range (Chernyshenko & Ones, 1999). This methodological problem is considered in two recent meta-analyses (Kuncel et al., 2007; Kuncel et al., 2001). In both studies the GRE subject tests were better predictors of cumulative grades for the first and the last PhD year than the GRE reasoning subtests. These findings are valid for most majors in general and for Social sciences in particular (Kuncel et al., 2001).

Results from the reported validation studies reveal that the usage of domain knowledge tests in student admission in addition to high school GPA can contribute to the prediction of study success. Another reason for their application is the lack of comparability of the high school GPA among different federal states in Germany. The same high school GPA is associated with different performance in the related knowledge tests depending on the attended school and the federal state (Köller, Baumert, & Schnabel, 1999). The legislative requirement to use high school GPA as the major selection criterion in college admission in Germany has led to the discrimination of applicants from federal states with higher educational school requirements or stronger grading. Attempts to counter this problem were the introduction of different weights for high school GPA from different federal states and the implementation of centrally developed and administered final examinations. The additional use of knowledge tests in student selection procedures could farther compensate the lack of GPA

comparability, since these tests provide an objective measure of knowledge, thus ensuring that applicants meet the requirements for a course they want to attend.

## Problems with measuring of study success

College grades undoubtedly represent a very important measure of study success. However, they are often used not as an operationalization, but as a substitute for study success. This assumption disregards three serious problems that are associated with college grades – namely their variance restriction, dubious reliability, and dimensionality.

Schneller and Schneider (2005) reported that in only 6 out of 34 German universities, the standard deviation for the cumulative graduate grade in Psychology was higher than $SD$ = .50 leading to a smaller variance than expected given the range of college grades.

Furthermore, results from several studies indicate that the objectivity and reliability of grades are rather low. For Medical sciences it was shown that the consensus between two independent judges rating the same examinee about the same topic with only a minor time lag ranges between $r$ = .40 and $r$ = .60 (Ingenkamp, 1975; Birkel, 1987). The correlation between oral and written examinations varies between $r$ = 0 and $r$ = .70, the majority of correlations varying between $r$ = .20 and $r$ = .40 (Ingenkamp, 1975; Birkel, 1987).

Pritz (1981) explored possible error sources for the low reliability of oral exams. In his study, students with higher oral fluency were given better grades after controlling for gesture and content. Oral fluency explained 17 % of the variance between the rater judgements. The variation of relevant previous grades explained another 7 % of this variance.

There are numerous other effects mainly investigated in Social Psychology which influence the result of an oral examination. In general, oral examinations can be biased by memory inferences, judgement preferences of the examinant (e.g. "halo" effect, tendency to the middle), position effects (primacy or recency effects), or self-fulfilling prophecies. While some of these effects can appear in written examinations as well, in those standardized situations the impact of social and verbal abilities is notably reduced.

Another aspect we want to address in the present article is the dimensionality of college grades. In educational research it is usually assumed that single college grades refer to one latent dimension of “study success”. Integrating all grades into one indicator GPA assumes that the available information across all grades is exhausted. However, this assumption has rarely been explicitly tested so far and first analyses indicate that it does not always hold true (Jonkmann, Wilhelm, & Leser, 2005). In that study, a separation between theoretical and mathematical classes on the one hand and applied and practical classes on the other hand seemed to be appropriate for available college grades in German undergraduate studies in Computer Science.

Furthermore, it remains often unclear whether grades are based upon a norm or upon a criterion. In general, they are assumed to be based upon a criterion meaning that grades should indicate to which degree a criterion was achieved. However, examination grades are often based upon a mixture between a norm and criterion. Consequently, lack of comparability is often associated with the grades because evaluation standards vary between different classes, schools, and universities. This problem is inherent in most studies analysing the prediction of study success.

## Studies overview

We have conducted three studies to investigate three different aspects of student admission that have not been appropriately investigated so far. In Study 1, we investigated the structure of college grades, estimating competing latent factors models for exam grades. In Study 2, we operationalized "study success" by taking into account undergraduate college grades as well as acquired knowledge in Psychology, measured through a curricular valid Psychology knowledge test. In Study 3, we strived to separate fluid and crystallized intelligence, conceptualized as domain specific knowledge, in student admission for German Psychology departments. We have explored the incremental and prognostic validity of these measures and high school GPA in predicting study success.

## Study 1

### *Objectives*

In this study we have explored the structure of college grades in Psychology by testing competing latent factors models for these grades. First, we tested a g-factor model which implies that a single latent variable is sufficient to account for the covariation between all college grades. In the second model, we distinguished between grades of basic, methodological classes versus grades of applied, practical classes. The third model distinguished between undergraduate and graduate college grades. We are not aware of any attempt to test different assumptions explicitly with a confirmatory measurement model and hope to shed some light on the grades' structure by our analyses. A distinction between oral and written exams was not possible for the available data due to different exam regulations for different student cohorts.

### *Methods*

We were granted access to all electronically registered university files for 910 Psychology students who started their studies at the Humboldt-Universität zu Berlin between 1990 and 2003. Files with complete undergraduate and graduate grades were available for 629 alumni (69%), who had successfully graduated from the Humboldt-Universität zu Berlin between 1996 and 2007. The reasons for incomplete datasets were drop-out, changing universities, or the fact that Psychology graduate studies were not completed yet. The German undergraduate diploma lasts about six to twelve months less than a Bachelor course. After completing their undergraduate degree, students continue with their graduate diploma that takes about another three and a half years. A Diploma degree is comparable to a Masters degree.

The current German undergraduate degree in Psychology covers the disciplines of Cognitive Psychology, Physiological Psychology, Developmental Psychology, Differential Psychology, Social Psychology, and Statistical Methods. Graduate diploma studies in Psychology include the disciplines of Clinical Psychology, Industrial and Organizational Psychology, Educational Psychology, Psychological Assessment, and Advanced Statistical Methods.

These subjects are taught with comparable curricula and comparable length in all German Psychology departments. In the graduate program students additionally specialize in one psychological discipline and choose one non-psychological subject. Grades in these latter subjects were not taken into account in the current study because they were not comparable between different student cohorts.

College grades in Germany range from "very good" (1) to "insufficient" (5). A subject is only completed when a grade of 4 or better has been achieved. Therefore when data of graduated students is analyzed, grades only range between 1 and 4 comprising 10 increments (1.0, 1.3, 1.7, 2.0, …, 3.7, 4.0).

## *Results*

### *Descriptive statistics*

Descriptive statistics for all grades are summarized in Table 1. The mean values varied slightly between different psychological disciplines. Undergraduate students performed best in Developmental Psychology, while graduate students excelled in Educational Psychology. Statistical Methods appeared to be the most challenging course in both undergraduate and graduate programs. The mean grade in undergraduate studies ($M$ = 2.35) was significantly higher than the mean grade in graduate studies ($M$ = 1.97; $t$ = 21.6; $p$ <. 001). Several reasons can account for this finding. We assume that it is not very likely that evaluation standards are less demanding in the graduate program, but rather that students might show more effort in later stages of the course because the grades they earn in the graduate program will play an important role for future job applications. It is also possible that these results indicate a stronger alignment between the students' interests and the contents offered in the graduate

**Table 1:**
Descriptive statistics for the college grades in study 1 (n = 629)

| | M | SE | SD | Min | Max |
|---|---|---|---|---|---|
| **Undergraduate degree** | | | | | |
| Cognitive Psychology | 2.3 | .03 | .71 | 1.0 | 4.0 |
| Physiological Psychology | 2.8 | .03 | .82 | 1.0 | 4.0 |
| Developmental Psychology | 1.9 | .03 | .77 | 1.0 | 4.0 |
| Differential Psychology | 2.2 | .03 | .87 | 1.0 | 4.0 |
| Social Psychology | 2.1 | .03 | .81 | 1.0 | 4.0 |
| Statistical Methods | 2.8 | .03 | .82 | 1.0 | 4.0 |
| Undergraduate GPA | 2.4 | .02 | .59 | 1.0 | 3.8 |
| **Graduate degree** | | | | | |
| Clinical Psychology | 2.0 | .03 | .79 | 1.0 | 4.0 |
| Industrial and Economical Psychology | 1.8 | .03 | .71 | 1.0 | 4.0 |
| Educational Psychology | 1.7 | .02 | .51 | 1.0 | 3.3 |
| Psychological Assessment | 2.1 | .02 | .58 | 1.0 | 4.0 |
| Advanced Statistical Methods | 2.3 | .03 | .82 | 1.0 | 4.0 |
| Graduate GPA | 2.0 | .02 | .50 | 1.0 | 3.6 |

*Notes:* M - mean, SE - standard error, SD - standard deviation, min - minimum, max - maximum
The column names M, SE und SD should be italicized.

program as compared to the undergraduate one. Another reason might be the fact that graduate classes are less crowded allowing for better student mentoring. Cronbach's alpha was $\alpha$ = .83 for undergraduate and $\alpha$ = .75 for graduate grades.

*Structural equation models*

We used confirmatory factor analysis techniques within the structural equation modelling framework (Kaplan, 2000; Bollen & Long, 1993; Bollen, 1989) to investigate the underlying structure of the college grades. For the estimation of this and all further reported measurement and structural models the statistical program Mplus 4.2 was used (Muthén & Muthén, 1998).

The model fit indices for the competing models described above are presented in Table 2. The g-factor model (Model 1) had an acceptable fit according to criteria by Hu and Bentler (1999): the Comparative Fit Index (CFI) exceeded .95 and the Root Means Square Error of Approximation (RMSEA) was less than .05. Model 2 which assumed two correlated factors for grades from basic classes vs. grades from applied classes yielded nearly identical model fit indices as Model 1. However, the residual covariance matrix was not positively definite for this model due to the fact that the estimated correlation between the two latent factors was $r$ = 1, implying that these factors are undistinguishable. Model 3 fitted the data excellently (see also Figure 1). The correlation between the latent factors was $r$ = .84 which indicated that both latent variables share a high amount of variance but are not identical. The Likelihood Ratio Test indicated that Model 3 fits the data significantly better than Model 1 ($\chi^2$ = 81; $df$ = 1; $p$ < .0001).

## *Discussion*

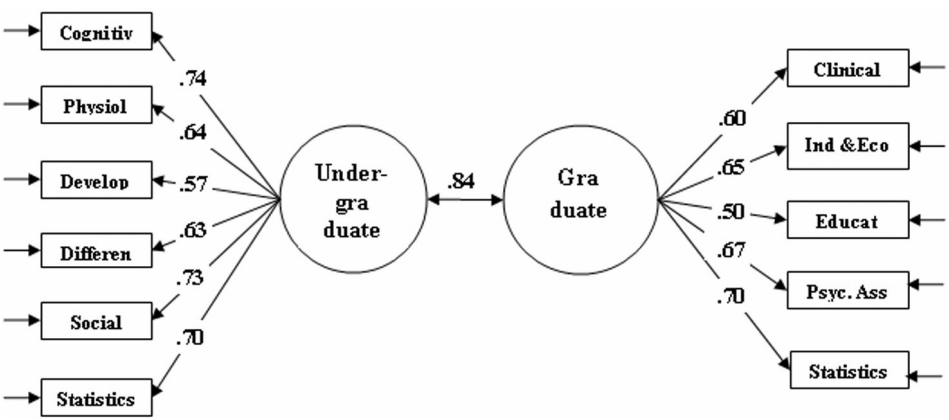
In the presented study a measurement model which accounted for a distinction between undergraduate and graduate grades fitted the data very well. The latent correlation between the two factors was high but reliably different from one. We assume that undergraduate and graduate grades will show different correlations with other measures. More specifically, we think it is possible that on top of the set of established predictors for undergraduate GPA, non-cognitive factors might proof incrementally valid in a performance context that provides little guidance and direction as a graduate course does.

The distinction between undergraduate and graduate grades in the final measurement model implies that the factor time can play an important role. Many studies tackling the issue of validity rely on undergraduate grades as the criterion of study success. In contrast, graduate grades are often not assessed mainly due to practical limitations. However, if undergraduate and graduate grades cannot be ascribed to the same underlying latent dimension, conclusions derived from undergraduate grades might not hold true for graduate grades. Our results show that the structure of college grades cannot always be assumed to be one-dimensional and suggest that their dimensionality should be tested explicitly when they are used as indicators of study success.

**Table 2:**
Model Fit Indices for the estimated measurement models for Psychology grades (n = 629)

| | Model Description | $\chi^2$ | *df* | *p* | *CFI* | *RMSEA* |
|---|---|---|---|---|---|---|
| 1 | g-factor model | 137 | 44 | < .001 | .958 | .06 |
| 2 | two correlated factors - basic vs. applied courses | 137 | 43 | < .001 | .957 | .06 |
| 3 | two correlated factors - undergraduate GPA vs. graduate GPA | 56 | 43 | .087 | .994 | .02 |

*Notes: df* – degrees of freedom; *CFI* - Comparative Fit Index; *RMSEA* - Root Mean Square Error of Approximation




*Notes:* $\chi^2 = 56$; $df = 43$; $p = .087$; *CFI* = .994; *RMSEA* = .022
*df* – degrees of freedom; *CFI* - Comparative Fit Index; *RMSEA* - Root Mean Square Error of Approximation
All path coefficients are statistically significantly different from zero at $\alpha = .05$

**Figure 1:**
Final measurement model for the college grades in Study 1 (n = 629)

## Study 2

### *Objectives*

In Study 2 we intended to widen the criterion space and to explore the structure of "study success" by defining it via two different methods for expressing student performance: undergraduate college grades in Psychology and an abbreviated German version of the Graduate Record Examination (GRE) subject test for Psychology (ETS, 2001). We expected to find a high but not perfect overlap of these two measures with an additional nested factor to account for common method variance.

The inspection of items in a version of the GRE Psychology test revealed that they highly overlap with topics covered in the German curriculum for undergraduate Psychology studies. The main difference between the German and American program is the coverage of Clinical Psychology. This discipline is only taught in the graduate program in Germany. For this reason items covering Clinical Psychology were not considered in the abbreviated German version of the test.

We translated a selection of items from the GRE Psychology into German and compiled a test version that mirrors the weighting of the different disciplines according to the German curriculum. Items from the following domains were included: Cognitive, Social, Physiological, Differential and Developmental Psychology, and Statistical Methods with emphasis on items covering Cognitive Psychology and Statistical Methods due to their relevance in the German curriculum. The final version of the test included 50 multiple-choice items with five response alternatives each, including 5 easy warming-up questions which were omitted from computations of the total score.

A variety of competing measurement models for the Psychology knowledge test and college grades were tested. First, a general factor measurement model for the Psychology knowledge test was estimated. Due to theoretical considerations the g-factor-model was modified by adding a correlated error between the summed scores for Physiological and Cognitive Psychology because both disciplines have a strong experimental orientation. In a third measurement model, a g-factor-model for the undergraduate grades was established.

Furthermore, four competing structural models were estimated by integrating the measurement models for the Psychology knowledge test and undergraduate grades. In the first structural model, a g-factor model for undergraduate GPA and Psychology knowledge test was calculated (Model 4). In the second structural model, two correlated latent factors for undergraduate GPA and Psychology knowledge test were assumed (Model 5). In the third model, a g-factor model with a nested factor that captures the grades' common method variance was tested (Model 6). Since the measurement model for the Psychology knowledge test implied the existence of common variance between the sub-scores for Physiological and Cognitive Psychology, this model was modified by introducing a second nested factor "Experimental orientation" to account for interest in experimental Psychology (Model 7).

## *Methods*

### *Sample*

Data was collected from November 2004 till January 2005. 183 students (139 female and 34 male) from five different German universities (24% Bielefeld, 23.5% Berlin, 8.2% Eichstätt, 10.9% Greifswald, 33.3% Trier) who had completed their undergraduate studies participated in the study. For ten participants demographic information was not available. The mean age was $M = 24.03$ years ($SD = 3.07$ years; ranging from 20 to 38 years). The average number of semesters completed at the time of investigation was $M = 8$ semesters ($SD = 1.7$), ranging between 5 and 19 semesters. The selected universities are not necessarily representative for Germany since their selection was based on existing cooperation with other German universities and was not randomly sampled. The students were recruited by their professors during their lectures. Participation in the study was voluntary.

All students reported their high school GPA and their exam grades in the six disciplines: Cognitive, Physiological, Developmental, Differential and Social Psychology, and Statistical Methods. For approximately 80 % of the students the self-reported grades could be validated by comparing them to the official university records. The correlations between self-reported and officially confirmed grades ranged between $r$ = .93 and $r$ = .99 across all disciplines. This finding supports the trustworthiness of the provided self-reported grades.

Four participants who completed less than 30 of 45 items in the Psychology knowledge test were removed from all further analyses. Missing values ranged between 0% and 16.8% for items ($M$ = 4.0%, $SD$ = 5%) and varied from 0% to 33% for students ($M$ = 4%, $SD$ = 0.8%). In total approximately 4% of all answers were missing. To test whether the number of missing values was a function of the sequence or position of an item in the test we calculated the number of correct answers in the first and second test-halves of the Psychology knowledge test (items 1-23 vs. 24-45). Since the T-Test for dependend samples was not significant ($t$ = 1.315, $df$ = 178, $p$ = 0.190) missing answers were replaced by 0 for all remaining 179 cases, based on the assumption that subjects skipped the items because they did not know the correct answer.

## *Results*

### *Descriptive statistics*

Item and scale analysis led to the exclusion of 10 items because these items were either too easy (item difficulty - proportion correct - higher than $P$ = .90) or too hard (item difficulty smaller than $P$ = .10). The final test version included 35 items, containing 8 items for Cognitive Psychology, 4 items for Social Psychology, 7 items for Physiological Psychology, 5 items for Differential Personality, 4 items for Developmental Psychology, and 7 items for Statistical Methods. After the item selection procedure the items' difficulties ranged between $P$ = .31 and $P$ = .86 with a mean of $P$ = .64. Although the Psychology knowledge test was fairly easy, no ceiling effects were observed. The mean total score in the final test version was $M$ = 22.3 ($SD$ = 4.7, ranging between 9 and 34). Cronbach's alpha was $\alpha$ = .69 with a mean inter-item-correlation of $r_{ij}$ = .06.

The internal consistency is rather moderate perhaps due to the following two reasons. Coefficient alpha is a function of the number of items, the mean inter-item-correlation (covariance) and item redundancy (Cortina, 1993). Low internal consistency can also be due to heterogeneity of the measure. Furthermore, it is possible that the internal consistency of a test is underestimated when the calculation is based on Pearson-product-moment-correlations. The results of the reliability analysis based on tetrachoric correlations provide much higher values for Cronbach's alpha ($\alpha$ = .81) and mean inter-item-correlation $r_{ij}$ = .12. Pearson-product-moment-correlations lead to an underestimation of Cronbach's alpha whereas tetrachoric correlations tend to overestimate it (Nunally, 1970).

At least five undergraduate grades were available for all students. Missing high school GPA (6 %) and exam grades (0.5 – 1.1 %) were imputed using the EM-Algorithm. The MCAR Little-Test (Little, 1988) was not significant ($\chi^2$ = 90.87; $df$ = 427; $p > .999$), implying that the data is missing completely at random and that its absence is not a function of other observed or unobserved variables.

*Measurement and Structural equation models*

Due to the rather small number of observations for the Psychology knowledge test in relation to the number of items, the complexity of the measurement models was reduced by using the sum scores for the six Psychology disciplines as manifest variables instead of the original binary answers. Model fit indices for all tested models described above are presented in Table 3. Model fit indices yielded an acceptable fit for the general factor model for the Psychology knowledge test (Model 1). The introduction of correlated errors in Model 2 led to an improvement in all fit indices.

Whereas the CFI of the g-factor-model for the undergraduate GPA (Model 3) was acceptable, the RMSEA value indicated misfit of the model. One plausible explanation for the high RMSEA value are the high zero-order correlations between the college grades – mean $r_{ij}$ = .42 in Model 3 vs. mean $r_{ij}$ = .25 in Models 1 and 2 and mean $r_{ij}$ = .27 in Models 4 to 7 (Browne, MacCallum, Kim, Andersen, & Glaser, 2002). The model with correlated errors for grades in Physiological and Cognitive Psychology had almost the same fit as Model 3. In this model the correlated errors were not statistically significant different from zero.

Model fit indices yielded an unacceptable fit for the first structural model (g-factor model) for undergraduate GPA and Psychology knowledge test (Model 4). All fit indices improved when a model with two distinct factors for grades and the test was established (Model 5). The Likelihood Ratio Test indicated that Model 5 fits the data significantly better than Model 4 ($\chi^2 = 43$; $df = 1$; $p < .0001$). The high correlation of $r = -.65$ between the latent

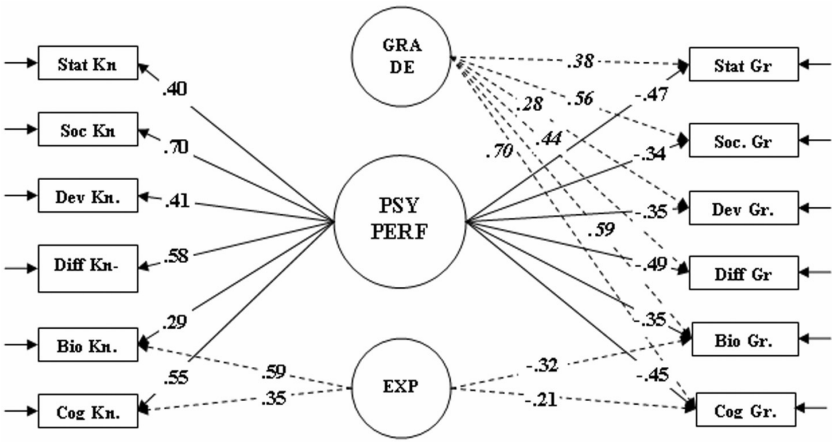
**Table 3:**

Model Fit Indices for the estimated measurement and structural models for the Psychology Knowledge Test and Psychology exam grades (n = 179)

| | | **Model Description** | ***χ²*** | ***df*** | ***p*** | ***CFI*** | ***RMSEA*** |
|---|---|---|---|---|---|---|---|
| 1 | Psychology Knowledge | g-factor model | 13.1 | 9 | .156 | .968 | .05 |
| 2 | Psychology Knowledge | g-factor model with correlated errors between Physiological and Cognitive Psychology | 5.0 | 8 | .751 | 1 | 0 |
| 3 | undergraduate GPA | g-factor model | 21.6 | 9 | .010 | .961 | .09 |
| 4 | Structural model | g-factor model for grades and knowledge | 129.2 | 54 | <.001 | .853 | .09 |
| 5 | Structural model | two correlated factors for grades and knowledge | 86.2 | 53 | .003 | .935 | .06 |
| 6 | Structural model | g-factor model with nested method factor for grades | 79.7 | 48 | .003 | .938 | .06 |
| 7 | Structural model | g-factor model with nested method factor for grades and nested factor for experimental orientation | 58.8 | 44 | .067 | .971 | .04 |

*Notes: df* – degrees of freedom, *CFI* - Comparative Fit Index, *RMSEA* - Root Mean Square Error of Approximation.

factors for undergraduate GPA and Psychology knowledge test in Model 5 indicated a high overlap between the constructs; however, they were not the same. For Models 5 and 6 CFI and RMSEA values indicated acceptable model fit. All factor loadings from the nested factor in Model 5 were substantially different from zero, implying that a substantive method factor for grades could be established. Model fit indices for Model 7 with a second nested factor "Experimental orientation" (see also Figure 2) indicated that it fitted the data very well. Note that factor loadings from the general factor to college grades were negative since the grades are reverse scaled, a low number indicating better performance and vice versa.




*Notes:* $\chi^2$ = 58.8; *df* = 44; *p* = .067; *CFI* = .971; *RMSEA* = .04
*df* – degrees of freedom; *CFI* - Comparative Fit Index; *RMSEA* - Root Mean Square Error of Approximation
All coefficients are statistically significant different from zero at $\alpha$ = .05
Loadings of the nested factors are pointed out by lines '-----'. The related factor loadings are italicized.

**Figure 2:**
Final structural model for Psychology grades and Psychology Knowledge Test in Study 2 (n = 179)

*Preliminary validity findings for the Psychology knowledge test*

For a subset of the sample ($n$ = 147 students) a portion of their graduate grades was available. The average number of available grades was $M$ = 4.8 ($SD$ = 1.96, ranging between 1 and 7). The average grade was $M$ = 1.64 ($SD$ = .51, ranging from 1.0 to 3.8).

We used the performance in the Psychology knowledge test for the prediction of graduate GPA to gather preliminary evidence on the validity of the Psychology knowledge test. The results are presented in Table 4. When interpreting these results take into account that the reliability of graduate GPA is smaller (Cronbach's alpha $\alpha$ = .43) than it was in Study 1 ($\alpha$ = .75) due to the lower number of available grades in the current study. High school GPA, undergraduate GPA and the total number of correct answers in the Psychology knowledge test were considered as predictors of graduate GPA. These predictors are correlated: $r$ = .36 for the correlation between high school GPA und undergraduate GPA, $r$ = -.51 between undergraduate GPA and Psychology knowledge test and $r$ = -.28 between high school GPA

**Table 4:**
Results of regression analyses for the prediction of graduate grade point average in German Psychology programmes (n = 147)

| Model | Predictor | *R* | $R^2$ | *corrected* $R^2$ | $\hat{\beta}$ | *t* | *p* |
|---|---|---|---|---|---|---|---|
| 1a | High school GPA | .25 | .06 | .06 | .25 | 3.10 | .002 |
| 1b | Undergraduate GPA | .30 | .09 | .09 | .30 | 3.83 | < .001 |
| 1c | Psychology knowledge | .29 | .09 | .08 | -.29 | -3.70 | < .001 |
| 2a | High school GPA | .34 | .11 | .10 | .16 | 1.92 | .057 |
| | Undergraduate GPA | | | | .25 | 2.92 | .004 |
| 2b | High school GPA | .34 | .12 | .10 | .18 | 2.22 | .028 |
| | Psychology knowledge | | | | -.24 | -2.98 | .003 |
| 2c | Undergraduate GPA | .34 | .12 | .11 | .21 | 2.27 | .025 |
| | Psychology knowledge | | | | -.19 | -2.06 | .042 |
| 3 | High school GPA | .37 | .14 | .12 | .14 | 1.70 | .091 |
| | Undergraduate GPA | | | | .17 | 1.76 | .080 |
| | Psychology knowledge | | | | -.17 | -1.85 | .066 |

*Notes:* GPA – grade point average, *R* – multiple regression coefficient, $R^2$ – coefficient of determination, $\hat{\beta}$ – standardized beta coefficient

and Psychology knowledge test. These relations indicate that collinearity effects for the three measures might be present, leading to possible problems when conducting multiple regression analysis. However, since the correlations between the measures are not high but only moderate, collinearity effects should not lead to major distortions in our results. All three predictors significantly contribute to the explanation of graduate GPA, either being considered alone or in combination with another predictor. Nevertheless, collinearity effects probably led to the non significant standardized regression coefficients when all three measures were considered simultaneously.

### *Discussion*

The results of this study show that grades and the curricular valid standardized test did not measure exactly the same aspects of study success. The final structural equation model implied that in addition to a performance factor, a nested method factor existed. Therefore, the criterion "study success" should be assessed via different measures in order to capture it more exhaustively.

Another important finding was that Psychology knowledge test was incrementally valid in the prediction of the graduate GPA after high school GPA or undergraduate GPA were considered. The observed regression coefficients for all measures in predicting graduate GPA were rather small but comparable to uncorrected coefficients as reported in GRE validity studies (Goldberg et al., 1992). If correction for the unreliability of the criterion was

taken into account, these coefficients would necessarily rise. Obviously, these preliminary validity results should be evaluated with caution given the presence of collinearity effects and the limited sample size. Nevertheless the results indicate that the application of standardized knowledge tests can substantially contribute to the prediction of study success in Master studies in addition to undergraduate college grades.

## Study 3

*Objectives*

In this study, we focussed on the research questions pertaining to the validity of fluid and crystallized intelligence, conceptualized as the domain specific knowledge. The main goal was to investigate the predictive and incremental validity of the domain knowledge tests when high school GPA and fluid intelligence measures were taken into account.

Job analysis indicate that knowledge in three disciplines is an important prerequisite in order to attain a degree in Psychology: English, Mathematics, and Biology (Wetzenstein, 2004). Proficiency in English is an important precondition for a successful graduation since the scientific literature is predominantly written in English. Basic knowledge in Mathematics is a relevant requirement because drop-out in Psychology is primarily caused by failures in Statistics and Quantitative Methods modules. Prior knowledge in selected areas of Biology was expected to be especially relevant for the success in disciplines with a substantial physiological focus.

Since we could not find any standardized tests for English, Mathematics, and Biology that would be suitable for our study, we have developed a specific knowledge test for each of the three domains. The English test was developed in cooperation with the Center for Foreign Languages department of the Humboldt-Universität zu Berlin. Items were designed to assess knowledge in English as a foreign language with respect to grammar, vocabulary, and complex test comprehension. Most items comprised of one sentence with one word missing. The subjects had to choose one correct answer among four alternatives. In order to discriminate well between persons with high ability levels, six items from the SAT I preparation book (College Board, 2003) were added. These items were hypothesized to be more difficult because two words were missing in each sentence. The original test included 40 items.

The Mathematics Test was constructed to measure applicant's skills in Algebra and Statistics. Algebra items were developed in accordance to existing items of the Mathematics test for high-school graduates and freshmen, the M-T-A-S (Lienert, Hofer, & Beleites, 1972). Items assessing basic knowledge in Statistics were developed by the authors of this paper, covering topics from the high-school curriculum. A multiple-choice answer format with four alternatives was chosen in order to reduce the guessing probability. The original version of the Mathematics test consisted of 21 items.

The Biology test was also developed by the authors of this paper. To ensure that the test has high content validity, items were developed according to the high-school curriculum for Biology focusing on topics taught in a standard biology book for high-school graduates and which at the same time are relevant to Psychology, namely physiology (metabolism, neurophysiology) and human biology (e. g. anatomy). The original version of the test contained 28 items. Table 5 shows sample items for each of the domain specific knowledge tests.

**Table 5:**
Examples of items from the domain specific knowledge tests

| Test | Item |
|---|---|
| English Test | The research is so ……. that it leaves no part of the issue unexamined.<br>*a) comprehensive*<br>b) rewarding<br>c) sporadic<br>d) problematical |
| Mathematics Test | 3 (x + 2b) = 15x + 6b<br>a) x = b<br>b) x = -b<br>*c) x = 0*<br>d) x = 1/b |
| Biology Test | How high is the resting membrane potential in a neurone?<br>a) -20mV<br>b) +60 bis +90 mV<br>c) +20 mV<br>*d) -60 bis -90 mV* |

*Notes:* Items in the Mathematics and Biology were in German in the original tests.
Correct responses are italicized.

We established a g-factor model for the Mathematics Test, assuming that one latent trait underlies performance in all test items. For the Biology test, a g-factor model was established as well. Because three items were related to the same topic – neurophysiologic processes in the cell membrane – the error terms of these items were correlated in the model.

For the English Test, two competing measurement models were estimated. The first tested model was a g-factor model. Because different aspects of English language skills were assessed by the test, the second model that included two nested factors in addition to the g-factor was established. One of these nested factors accounted for common variance among items assessing grammar skills, while the other accounted for variance among items with a high demand on text comprehension.

For fluid intelligence, a nested factor measurement model was established. It comprised one general factor representing "fluid intelligence" and three additional nested factors which were orthogonal to one another and to the g-factor. The nested factor for "verbal ability" represented the ability to deal with verbal material. The second nested factor for "numerical ability" represented the ability to deal with numeric material. The third nested factor, "General Speed" represented a speed component in the according tasks when the impact of fluid intelligence has been statistically controlled.

## *Methods*

### *Sample*

Data was collected from November till December 2004. 387 undergraduate students (305 female and 74 male) in their first college semester from five different German universities (17.6% Berlin, 20.3% Bielefeld, 14.6% Greifswald, 12.5% Potsdam, 34.9% Trier) were tested. Participants ranged in age from 18 to 49 ($M$ = 22 years, $SD$ = 3.8 years). Eight persons did not give particulars on their sex, 10 other participants did not specify their age. All students were asked to report their high school GPA and to work on a large test battery. As in study 2, the sampled universities were not chosen randomly but their participation was based on prior cooperation. However, there is no obvious sampling bias. Students were recruited by the teaching professors during their lectures. Participation in the study was voluntary.

Two years after taking the tests, undergraduate grades in Cognitive, Physiological, Developmental, Differential and Social Psychology, and Statistical Methods were collected from the university records. Data was available for 295 students. The number of available exam grades ranged from 1 to 6 ($M$ = 5.4; $SD$ = 1.2). Students with a small number of exam grades available were not excluded from further analysis since they compose a group that has not been successful so far - these students have either already dropped out of college or might do so in the future. Retention is another possible criterion of study success. Even if these students did not drop out of the college, they are less successful than other students who have already finished their undergraduate studies, because they take longer to finish. Since we were interested in the sensitivity of the newly developed tests to detect unsuitable candidates, data of all 295 participants was retained for further analyses.

The popularity of the five selected universities differs, leading to a different selectivity of the student admission procedures (mainly based on high school GPA) in these universities. In order to reduce the institutional impact and subjective judging effects of individual examiners, grades for each Psychological discipline were standardized separately within each university.

### *Test battery*

In addition to the newly developed domain-specific knowledge tests, fluid intelligence and mental speed were assessed using 12 tasks from the Berlin Test of Intelligence Structure (Jäger et al., 2006). In his model of intelligence Jäger (1982, 1984) assumes two independent dimensions on which tasks which assess intelligence can vary - operations and content. Each task measures one of the four operations (fluid intelligence, mental speed, memory, creativity) with one of the three content facets (verbal, numerical, figural).

We have compiled a version of the test by choosing four numerical, four verbal and four figural tasks, half of them assessing fluid intelligence (e.g. verbal and figural analogies, number and figure series), the other half assessing mental speed using easy tasks with severe time limits (e.g. finding and crossing letters, odd numbers or words belonging to certain categories).

## *Results*

### *Data analyses*

Missing values for the majority of knowledge items ranged between 0 % and 11 %. Exceptions were the English items which were taken from the SAT preparation book. For these items up to 24 % of the answers were missing. These items require very advanced proficiency in English. Hence, we assume that participants did not provide an answer because they did not know it. Therefore, missing answers for all knowledge items (4.4% for the English test, 3.9% for the Mathematics test and less than 1% for the Biology test) were replaced by 0, implying that subjects skipped the items because of the lack of knowledge.

Missing high school GPA (4.4 %) and missing scores for particular fluid intelligence tasks (0.3 %) were imputed using the expectation maximisation (EM) algorithm. The MCAR Little-Test yielded no significant result ($\chi^2 = 95.7$; $df = 131$; $p = .991$), implying that the data is missing completely at random.

In further data analyses items were retained which fulfilled the following criteria: a) item difficulty was $.1 < P < .9$ and b) corrected item-test correlation was $r_{it} > .1$. These criteria were applied to all three knowledge tests. Following these criteria, six items were eliminated in the English knowledge tests, four items in the Mathematic knowledge test and 12 items in the Biology knowledge test. Two items with difficulties higher than $P > .9$ were not excluded from further analyses due to their high item-test correlation.

### *Descriptive statistics*

Means, standard errors, standard deviations, item difficulties, corrected item-test correlations, Cronbach's alpha and reliability of the latent factor ω for all knowledge tests are presented in Table 6. Item difficulties range between $P = .20$ and $P = .94$, indicating that easy, average and difficult items are contained in each test. Although all three knowledge tests were fairly easy (see Table 6) no ceiling effects were observed.

Cronbach's alpha is high for the English knowledge test ($\alpha = .84$), while it is only acceptable for the Mathematics and Biology tests ($\alpha = .60$ and $\alpha = .62$). As argued before (see Study 2), low internal consistency can be due to a small number of items in a test or to the heterogeneity of the measures of interest. The values for Cronbach's alpha for these measures rise when their calculation is based on tetrachoric correlations instead of Pearson moment product correlations (see Table 6). Additionally, the reliability of the latent factor ω was computed through confirmatory factor analysis specifying a single latent factor for each test based upon all items included in this test (McDonald, 1999). The comparison of all three reliability values enables the appraisal of the under- and overestimation of the reliability for Pearson and tetrachoric correlation matrices, respectively. Values of Cronbach's alpha based on tetrachoric correlations are very close to those for the latent reliability ω, while Cronbach's alpha relying on Pearson correlations is on average .12 smaller than ω, indicating a considerable underestimation of the internal consistency by the Pearson-product-moment correlations. Furthermore, the mean tetrachoric inter-item correlation is provided for each knowledge test.

**Table 6:**
Descriptive statistics for the knowledge tests (n = 387)

| Test | *n* | *Mean* | *SD* | *P* | *mean P* | $r_{it\_bis}$ | *mean* $r_{ij\_tetra}$ | $\alpha$ | $\alpha_{tetra}$ | $\omega$ |
|---|---|---|---|---|---|---|---|---|---|---|
| English Test | 36 | 20.29 | 6.12 | .23 - .93 | .60 | .24 - .72 | .28 | .84 | .92 | .92 |
| Mathematics Test | 17 | 9.77 | 2.86 | .27 - .81 | .57 | .14 - .50 | .15 | .60 | .74 | .74 |
| Biology Test | 16 | 10.08 | 2.70 | .20 - .94 | .63 | .22 - .66 | .21 | .62 | .79 | .77 |

*Notes: n* - number of items, *SD* – standard deviation, *P* - item difficulty, *mean P* – mean item difficulty, $r_{it\_bis}$ – corrected item-total correlation, *mean* $r_{ij\_tetra}$ – mean tetrachoric inter-item correlation, $\alpha$ - Cronbachs α based on Pearson product moment correlations, $\alpha_{tetra}$ - Cronbachs α based on tetrachoric correlations, $\omega$ - reliability of the latent factor

*Measurement Models*

Measurement models for the three knowledge tests were estimated using the robust WLSMV estimator (Weighted Least Standardized Means und Variance). It is the appropriate estimator for binary manifest variables (Muthén, 1984) and is implemented in Mplus 4.2. Models with continuous manifest variables were estimated using the Maximum-Likelihood-Algorithm. Analyses with the ML estimator were based on Pearson product moment correlations, while the parameter estimation with the WLSMV estimator was based on tetrachoric correlations.

Model fit indices for all tested models described above are presented in Table 7. The tested measurement models for the Mathematics and Biology knowledge tests (Models 1 & 2) fitted the data excellently. Model fit indices indicated misfit for the g-factor-model for the

**Table 7:**
Model Fit Indices for the estimated measurement models for the knowledge and fluid intelligence tests (n = 387)

| | Test | Model Description | $\chi^2$ | *df* | *p* | *CFI* | *RMSEA* |
|---|---|---|---|---|---|---|---|
| 1 | Mathematics Test | g-factor model | 79.0 | 84 | .634 | 1 | 0 |
| 2 | Biology Test | g-factor model with three correlated errors | 66.4 | 69 | .565 | 1 | 0 |
| 3 | English Test | g-factor model | 254.8 | 179 | < .001 | .940 | .03 |
| 4 | English Test | g-factor and two nested factors Grammar and Text comprehension | 196.6 | 177 | .149 | .984 | .02 |
| 5 | Fluid intelligence | g-factor and three nested factors Verbal, Numerical and General Speed | 48.8 | 23 | .001 | .969 | .05 |

*Notes: df* – degrees of freedom, *CFI* - Comparative Fit Index, *RMSEA* - Root Mean Square Error of Approximation

English knowledge test (Model 3). The introduction of the nested factors "grammar" and "text comprehension" in Model 4 led to better model fit. For further analyses, we accepted Model 4 as the final measurement model. Because of its partly explorative nature it needs to be replicated with independent data.

Model fit indices for the measurement model of fluid intelligence (Model 5) implied a very good fit. Based on these measurement models factor scores for all subjects on all latent factors were estimated. These scores were used in correlation and regression analyses.

*Correlation analyses*

Correlations between the factor scores for the three knowledge tests and the fluid intelligence test were calculated in order to investigate the relations between these measures. All correlations were substantially different from zero (see Table 8). The highest correlation for fluid intelligence was found with Mathematics ($r$ = .56), followed by the correlation with English ($r$ = .34), and last but not least with Biology ($r$ = .20). This confirms our hypothesis that performance in a Mathematics test particularly depends on fluid intelligence. The knowledge tests were moderately correlated among themselves with correlations ranging from $r$ = .16 (English and Mathematics) to $r$ = .28 (Mathematics and Biology).

The nested factors for general speed, grammar, and text comprehension were not considered in these analyses because no theoretical expectations existed for any relations to these factors. Numerical ability correlated highly with the factor score in the Mathematics test ($r$ = .37) as expected. At the same time, verbal ability had substantive relations with the factor scores for English ($r$ = .37) and, to a lesser extent, for Biology ($r$ = .26).

Strikingly, the factor scores for numerical ability, verbal ability, and fluid intelligence show statistically significant correlations among each other even though the latent factors were postulated to be orthogonal to one another in the measurement model. This result is due to the fact that factor scores were calculated using the regression method which does not preserve the model's structure. However, because factor scores for numerical and verbal ability were not considered in the following regression analysis, these statistically significant correlations do not compromise the results of the regression analyses.

**Table 8:**
Correlations between the factor scores for latent factors (n = 387)

| | Fluid Int. | Eng | Bio | Math | VER | NUM |
|---|---|---|---|---|---|---|
| Fluid intelligence (Fluid Int.) | *.73* | - | - | - | - | - |
| English knowledge (Eng) | **.34** | *.77* | - | - | - | - |
| Biology knowledge (Bio) | **.20** | **.21** | *.64* | - | - | - |
| Mathematic knowledge (Math) | **.56** | **.28** | **.16** | *.65* | - | - |
| Verbal ability (VER) | **.21** | **.37** | **.26** | .05 | *.43* | - |
| Numerical ability (NUM) | **.22** | .07 | -.05 | **.37** | **-.22** | *.46* |

*Notes:* Numbers in the diagonal show the variance of the factor scores for the latent factors. Numbers printed bold indicate that the particular correlation was significantly different from zero at $\alpha$ = .05.

*Regression analyses*

Hierarchical regression analysis was used to predict undergraduate GPA by high school GPA, fluid intelligence measures and crystallized intelligence operationalised via the knowledge tests. This approach enables us to estimate the incremental predictive validity of each predictor. Modelling the relations in a structural equation model was not possible due to the complexity of the measurement models and a small sample size of $n = 295$. Similarly, multigroup analysis to test the measurement invariance within the five different universities was not feasible because of small group sizes (37, 43, 52, 60 and 103 students, respectively). The results of the correlation analyses indicate collinearity effects for the measures of interest. As argued above (see Study 2) multiple regression analysis can be problematic when strong collinearity effects occur, but this is not necessarily the case with the moderate correlations found here. Indeed, in the present study collinearity effects do not seem to substantially compromise the results of the multiple regression analysis. Nevertheless the interpretation of the proportion of variance common to all performance measures analyzed here is an important issue. Simply labelling such a component "g" would not be adequate given the curricular nature of many of the indicators investigated here.

The main results for the regression analyses are provided in Table 9. Both high school GPA (Model 1b) and the knowledge tests (Model 1c) are substantially correlated with un-

**Table 9:**
Results of regression analyses for the prediction of undergraduate GPA (n = 295)

| Model | Predictors | *R* | $R^2$ | *corrected* $R^2$ | $\hat{\beta}$ | *t* | *p* |
|---|---|---|---|---|---|---|---|
| 1a | Fluid intelligence | .23 | .05 | .05 | -.23 | -4.04 | < .001 |
| 1b | High school GPA | .44 | .19 | .19 | .44 | 8.31 | < .001 |
| 1c | English | .44 | .19 | .18 | -.23 | -4.23 | < .001 |
| | Mathematics | | | | -.23 | -4.26 | < .001 |
| | Biology | | | | -.17 | -3.11 | .002 |
| 2a | Fluid intelligence | .46 | .21 | .20 | -.13 | -2.50 | .013 |
| | High school GPA | | | | .41 | 7.55 | < .001 |
| 2b | Fluid intelligence | .44 | .19 | .18 | -.01 | -.13 | .900 |
| | English | | | | -.23 | -4.15 | < .001 |
| | Mathematics | | | | -.23 | -3.56 | < .001 |
| | Biology | | | | -.17 | -3.08 | .002 |
| 2c | High school GPA | .52 | .27 | .26 | .32 | 5.79 | < .001 |
| | English | | | | -.16 | -2.89 | .004 |
| | Mathematics | | | | -.15 | -2.71 | .007 |
| | Biology | | | | -.14 | -2.78 | .006 |
| | Fluid intelligence | .52 | .27 | .26 | -.01 | -.08 | .933 |
| | High school GPA | | | | .32 | 5.77 | < .001 |
| 3 | English | | | | -.16 | -2.85 | .005 |
| | Mathematics | | | | -.14 | -2.31 | .021 |
| | Biology | | | | -.14 | -2.76 | .006 |

*Notes:* GPA – grade point average, R – multiple regression coefficient, $R^2$ – coefficient of determination, $\hat{\beta}$ – standardized beta coefficient

dergraduate GPA and predict nearly 19% of its variance, whereas fluid intelligence correlates only moderately with undergraduate GPA (Model 1a). The fact that the relation between college grades and fluid intelligence is weak renders the latter a small contributor of incremental variance over high school GPA or knowledge tests (Model 2a and 2b) in predicting undergraduate GPA. In contrast, the multiple correlation between high school GPA and knowledge tests on one side and undergraduate GPA on the other side is $r = .52$. Additional 8 % of the criterion variance can be explained when both knowledge tests and high school GPA are used as predictors in a regression model (Model 2c). This result indicates that specific knowledge tests are incrementally predictive over high school GPA in predicting exam grades.

The multiple correlation between all three predictors on one side and undergraduate GPA on the other side (Model 3) remains the same as it was in Model 2b when fluid intelligence was not included, indicating that there is no evidence for its incremental validity above high school GPA and knowledge tests. Possible reasons for this unexpected finding will be discussed in the next section.

## Discussion

In Study 3, confirmatory measurement models with satisfying model fit indices could be established for all ability tests. Knowledge tests were incrementally predictive over high school GPA in explaining variance of undergraduate college grades. This result suggests that using domain specific knowledge tests in a student admission process for German Psychology programs will increase the number of successful students. Surprisingly, fluid intelligence only had very small prognostic validity and was not incrementally valid over high school GPA and the knowledge tests. This result contradicts fundamental findings in intelligence and student admission research (Wilhelm & Engle, 2005; Kuncel et al, 2004).

Several possible reasons can explain this unexpected finding. Perhaps this is due to the specifics of the selective sample that was tested. All subjects in the study were students already enrolled at university. They were mainly selected on their high school GPA which is highly related to fluid intelligence (Ones et al., 2005; Snow & Yalow, 1982). Therefore, it is possible that the variance in fluid intelligence was strongly limited in the observed sample in comparison to representative national samples. Furthermore, participation in the study was not mandatory. It is possible that participants who would take a test voluntarily have higher abilities than average students. Both possible explanations constrain the variance in fluid intelligence which in turn would lead to an underestimation of its predictive validity.

In order to check the existence of restriction of range for fluid intelligence, we compared scores of our sample with data on the same test from an unselected sample of high-school students. The mean IQ of the 295 college students, based on standardized norms, was $M = 111.4$ ($SD = 10.8$), ranging between 84 and 136. In contrast, the mean IQ for the high school sample ($n = 329$) was significantly lower with $M = 102.5$ ($SD = 11.7$; ranging from 70 to 127; $t = 9.93$; $df = 622$; $p < .001$). Accordingly, the restriction of range in the student sample amounts to 17 %. Therefore, it can only partially explain the lack of predictive validity of fluid intelligence.

Moreover, the results imply that high school GPA had more predictive power than fluid intelligence in our data. Surprisingly, the restriction of range in high school GPA in our data

is smaller than expected. Since Psychology is a very popular subject, we expected the selection procedure based on high school GPA to lead to low variance in high school GPA. However, approximately 25 % of the students were admitted based on waiting time where the applicants are granted a bonus on high school GPA for every semester they have been waiting to enter the Psychology program. This obviously countered a restriction in variance successfully.

The results so far can not explain why fluid intelligence was not more predictive of college performance. Another possible reason is that performance in German Psychology programs might not depend as much on fluid intelligence as members of the psychological community are inclined to believe. Our prediction is that the predictive validity of fluid intelligence is contingent upon academic rigor and reliance on mathematics. We suggest that Psychology programs are more similar to social sciences and humanities than they are akin to natural sciences with regard to the contents and methodology taught. We predict that on top of the set of established predictors such as fluid intelligence, grades and knowledge noncognitive factors are more likely to be incrementally valid for social sciences and humanities than for natural sciences. Apparently, testing this hypothesis is an interesting topic for future research.

## General Discussion

In the first study the dimensionality of the college grades was examined. The comparison of the three competing models revealed that a distinction between undergraduate and graduate college grades seemed to be appropriate. In the second study we investigated the overlap between college grades and knowledge measures as the criteria of "study success". The high correlation between college grades and the Psychology knowledge test in a structural model indicates a high correlation between both measures; however, they are not the same. A substantive nested method factor accounting for common variance among grades can be established. We conclude that it's imperative to carefully choose an appropriate way of operationalising "study success" because selecting one specific measurement method has important implications on the results. In the third study, knowledge tests were found to be incrementally predictive over high school grades in explaining variance of undergraduate GPA. Surprisingly, fluid intelligence only had very small prognostic validity and was not incrementally valid over and above high school grades and knowledge tests.

These studies demonstrate some methodological problems that have not been investigated so far. They might have several implications for student admission procedures in the future. The results of Study 1 demonstrate that college grades from graduate and undergraduate studies refer to correlated but not identical latent dimensions. That is interesting and warrants further research because our results point to a limitation in much of the available evidence on the validity of admission indicators. The majority of validity studies only assess undergraduate GPA as a criterion (Schmidt-Atzert, 2005; Bridgeman et al., 2000; Ramist et al., 1993) and disregard potential differential relations with graduate GPA.

Findings in Study 2 suggest that an extension of the criterion definition of "study success" is required: Considering curricular valid standardized knowledge tests in addition to college grades can be appropriate in order to operationalize "study success" at the end of undergraduate studies. Within the Bologna process the German universities are undergoing a

variety of changes in curricula, implementing the international Bachelor and Master degrees. These changes might demand a revision of student admission procedures. More specifically, introduction of knowledge tests can provide a very useful instrument in student selection, particularly for the Master Degree. Although the validity of the GRE in the US is established beyond reasonable doubt, the use of standardized knowledge tests for admission into graduate programs has been a neglected opportunity so far.

The results of Study 3 suggest that the application of domain specific knowledge tests in a student admission process can enhance selection procedures. In spite of their indicated benefit, current juridical regulations can hinder their implementation in practice in Germany. However, the finding that the validity of fluid intelligence measures was only low in contrast to the validity of domain knowledge tests indicates that a distinction between fluid and crystallized intelligence deserves closer attention in the student selection context, too.

In the future we intend to collect complete graduate college grades from the participants in Study 3 in order to replicate and extend our results from Study 1 and operationalize "study success" more broadly by taking all grades into account. Moreover, these data will also allow us to assess retention at college as another relevant criterion of "study success". Obviously, tracking a larger sample would be helpful to investigate these research questions with more statistical power. It would be interesting to explore the relations between undergraduate GPA, the Psychology knowledge test, and the newly developed domain-specific knowledge tests in Study 3. Unfortunately, it was not possible to test the students with the Psychology knowledge test two years after the first test session due to privacy protection regulations.

Another promising tool which has recently been developed is a reading comprehension test specifically designed for the selection of Psychology students. It comprises texts, tables, and figures that cover psychological topics. A high level of fluid intelligence is associated with high performance in this "psychological science comprehension test" (Wilhelm et al., 2006; Formazin et al., in press).

Summarizing the main results, we strongly recommend a) to consider grades from undergraduate and graduate studies as well as results from standardized knowledge tests on the criterion side and b) to take fluid intelligence and relevant knowledge measures into account on the predictor side. More multivariate considerations on the predictor and criterion side in college admission problems can lead to a more appropriate use of available assessment tests (Formazin, Wilhelm & Köller, 2006).

## References

Birkel, P. (1978). *Mündliche Prüfungen. Zur Objektivität und Validität der Leistungsbeurteilung* [Oral exams. Objectivity and validity of performance evaluation]. Bochum: Kamp.

Bollen, K A (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K A & Long, S J. (1993). *Testing structural equation models*. London: SAGE Focus Edition.

Bridgeman, B., Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT I: Reasoning Test* (College Board Report No. 2000-1). New York: College Entrance Examination Board.

Bridgeman, B., Burton, N., & Cline, F. (2001*). Substituting SAT II: Subject tests for SAT I: Reasoning test: Impact on admitted class composition and quality* (College Board Research Report No. 2001-3). New York: College Entrance Examination Board.

Brody, N. (1992). *Intelligence.* San Diego, California: Academic Press.

Browne, M. W., MacCallum, R. C., Kim, C., Andersen, B. L., & Glaser, R. (2002). When fit indices and residuals are incompatible. *Psychological Methods*, 7, 403-421.

Burton, N.W. & Ramist, L. (2001). *Predicting success in college: SAT studies of classes graduating since 1980* (College Board Research Report No. 2000-2). New York: College Entrance Examination Board.

Burton, N.W. & Wang, M. (2005) *Predicting long-term success in graduate school: A collaborative validity study* (GRE Board Report Nr. 99-14R). Princeton: Educational Testing Service.

Camara, W. & Echternacht, G. (2000). *The SAT I and high school grades: Utility in predicting success in college* (College Board Research Notes 2000-10). New York: College Entrance Examination Board.

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

Cattell, R.B. (1987). *Intelligence: Its structure, growth and action.* Amsterdam: North-Holland.

Chernyshenko, O. S. & Ones, D. S. (1999). How selective are Psychology Graduate Programs? The effect of the selection ratio on GRE Score validity. *Educational and Psychological Measurement, 59,* 951 - 961.

College Board. (2003). *10 Real SATs* (3rd edition). New York: College Entrance Examination Board.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98–104.

Educational Testing Service [ETS]. (2001). *GRE Psychology Test - Practice Book.* Princeton: Educational Testing Service.

Fenster, A., Markus, K. A., Wiedemann, C. F., Brackett, M. A., & Fernandez, J. (2001). Selecting tomorrow's forensic psychologists: A fresh look at some familiar predictors. *Educational and Psychological Measurement, 61,* 336-348.

Formazin, M., Wilhelm, O., & Köller, O. (2006), Willkür vermeiden! Sachlich gebotene Methoden der Beurteilung von Studienbewerbern [Avoid arbitrariness! Appropriate methods in student admission]. *Forschung und Lehre, 13*, 672-674.

Formazin, M., Kunina, O., Schroeders, U., Wilhelm, O., Hildebrandt, A., & Köller, O. (in press). Validitäts- und Nützlichkeitsüberlegungen zur Studierendenauswahl im Allgemeinen mit Präzisierungen für das Fach Psychologie im Besonderen [Thoughts on validity and utility in student admission in general and for Psychology in particular]. In H. Schuler & B. Hell (Eds.), *Studierendenauswahl und Studienentscheidung*. Göttingen: Hogrefe.

Geiser, S. & Studley, R. (2002). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. *Educational Assessment, 8,* 1-26.

Gold, A. & Souvignier, E. (2005). Prognose der Studierfähigkeit: Ergebnisse aus Längsschnittanalysen [Predicting the ability to study: Results from longitudinal analyses]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 37,* 214-222.

Goldberg, E. L. & Alliger, G. M. (1992). Assessing the validity of the GRE for students in psychology: A validity generalization approach. *Educational and Psychological Measurement, 52,* 1019-1027.

Horn, J. L. (1988). Thinking about human abilities. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (pp. 645–685). New York: Plenum.

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

Ingenkamp, K. (1975). *Pädagogische Diagnostik. Ein Forschungsbericht über Schülerbeurteilung in Europa* [Educational assessment: Research report on evaluating students in Europe]. Weinheim: Beltz.

Jäger, A.O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen: Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells [Multidimensional classification of intelligence performance: Experimentally controlled development of a descriptive model of intelligence structure]. *Diagnostica, 28,*195-225.

Jäger, A.O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven [Intelligence structural research. Competing models, new developments, perspectives]. *Psychologische Rundschau*, *35,* 21-35.

Jäger, A.O., Holling, H., Preckel, F., Schulze, R., Vock, M., Süss, H.-M, & Beauducel, A. (2006). *Berliner Intelligenzstruktur Test für Jugendliche: Begabungs- und Hochbegabungsdiagnostik (BIS-HB)* [Berlin Test of Intelligence Structure for Teenagers: Assessment of gifted and highly gifted teenagers]. Göttingen: Hogrefe.

Jonkmann, K., Wilhelm, O., & Leser, U. (2005). *Studienabbruch, Studiendauer und Studienerleben: Analyse der Studierendenumfrage des Instituts für Informatik der Humboldt-Universität zu Berlin* [College drop-out, study duration and college experiences: Analysis of a survey amongst students at the Institute for Computer Sciences at the Humboldt University]. Unpublished report, Humboldt-Universität zu Berlin.

Kaplan, D. (2000). *Structural Equation Modelling: Foundations and extensions.* London: SAGE, Advanced Quantitative Techniques in the Social Sciences series.

Kobrin, J.L., Camara, W.J., & Milewski, G.B. (2002). *The utility of the SAT I and SAT II for admissions decisions in California and the nation* (Research Report No. 2002-6). New York: College Entrance Examination Board.

Köller, O., Baumert, J., & Schnabel, K. (1999). Wege zur Hochschulreife: Offenheit des Systems und Sicherung vergleichbarer Standards. Analysen am Beispiel der Mathematikleistungen von Oberstufenschülern an integrierten Gesamtschulen und Gymnasien in Nordrhein-Westfalen [Different ways to A-level: Openness of the system and securing comparable standards. Analyses for mathematical performance at comprehensive schools and high schools in North Rhine-Westphalia]. *Zeitschrift für Erziehungswissenschaft, 2,* 385-422.

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

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology*, *86,* 148-161.

Kuncel, N.R. & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science*, *315,* 1080 – 1081.

Lienert, G.A., Hofer, M., & Beleites, J. (1972). *Mathematiktest für Abiturienten und Studienanfänger (M-T-A-S)* [Mathematic test for high school students and freshman (M-T-A-S)]. Göttingen: Hogrefe.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, *83,* 1198-1202.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Morrison, T. & Morrison, M. (1995). A meta-analytic assessment of the predictive validity of the quantitative and verbal components of the Graduate Record Examination with graduate grade point average representing the criterion of graduate success. *Educational and Psychological Measurement, 55,* 309-316.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika, 49,* 115-132.

Muthén, L. K. & Muthén, B. O. (1998). *Mplus* (Version 4.2.) [Computer software]. Los Angeles, CA: Muthén & Muthén.

Noble, J. & Sawyer, R. (2002). *Predicting different levels of academic success in college using high school GPA and ACT Composite Score* (ACT Research Report Series 2002-4). Iowa City, IA: ACT, Inc.

Nunally, J. C. (1970). *Introduction to psychological measurement.* New York: McGraw-Hill.

Ones, D. S., Viswesvaran, C., & Dilchert, S. (2005). Cognitive ability in selection decisions. In: O. Wilhelm & R. W. Engle (Eds.), *Understanding and measuring intelligence* (pp. 431-461). London: Sage.

Pritz, V. (1981). Der Einfluss von Sprechflüssigkeit und Vorinformation auf die Leistungsbeurteilung in der mündlichen Reifeprüfung [The influence of oral fluency and previous information on evaluation of performance in oral A-level exam]. In K. Ingenkamp (Ed.), *Wert und Wirkung von Beurteilungsverfahren* (pp. 49-96). Weinheim: Beltz.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1993). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (College Board Report No. 93-1). New York: College Entrance Examination Board.

Ramist, L., Lewis, C., & McCamley-Jenkins, L. (2001). *Using achievement tests/SAT II Subject Tests to demonstrate achievement and predict college grades: Sex, language, ethnic, and parental education groups* (College Board Report No. 2001-5). New York: College Entrance Examination Board.

Schmidt-Atzert, L. (2005). Prädiktion von Studienerfolg bei Psychologiestudenten [Prediction of study success for Psychology students]. *Psychologische Rundschau, 56,* 131-133.

Schneller, K. & Schneider, W. (2005). Bundesweite Befragung der Absolventinnen und Absolventen des Jahres 2003 im Studiengang Psychologie [National survey of alumni in Psychology in 2003]. *Psychologische Rundschau, 56,* 159-175.

Schuler, H., Funke, U., & Baron-Boldt, J. (1990). Predictive validity of school grades: A meta-analysis. *Applied Psychology: An International Review*, *39,* 89-103.

Snow, R. E. & Yalow, E. (1982). Education and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 493-585). New York: Cambridge University Press.

Steyer, R., Yousfi, S., & Würfel, K. (2005). Prädiktion von Studienerfolg: Der Zusammenhang zwischen Schul- und Studiennoten im Diplomstudiengang Psychologie [Prediction of study success: Relations between high school and college grades in Psychology diploma studies]. *Psychologische Rundschau, 56,* 129-131.

Stumpf, H. & Nauels, H.-U. (1990). Zur prognostischen Validität des „Tests für medizinische Studiengänge" (TMS) im Studiengang Humanmedizin [Prognostic validity of the Medical Entrance Test TMS for human medicine]. *Diagnostica, 36,* 16-32.

Trapmann, S., Hell, B., Weigand, S., & Schuler, H. (in press). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs – eine Metaanalyse [Validity of high school grades in predicting study success – a meta analysis]. *Zeitschrift für Pädagogische Psychologie.*

Trost, G., Klieme, E., & Nauels, H.-U. (1997). Prognostische Validität des Tests für medizinische Studiengänge (TMS) [Prognostic validity of the Medical Entrance Test TMS]. In T. Hermann (Ed.), *Hochschulentwicklung – Aufgaben und Chancen* (pp. 57-78). Heidelberg: Asanger.

Trost, G., Blum, F., Fay, E., Klieme, E., Maichle, U., Meyer, M., & Nauels, H.-U. (1998). *Evaluation des Tests für Medizinische Studiengänge (TMS): Synopse der Ergebnisse* [Evaluation of the Medical Entrance Test TMS: Synopsis of the results]. ITB: Bonn.

Wetzenstein, E. (2004, April). *Entwicklung eines Anforderungsprofils für Studierende am Beispiel der Psychologie* [Development of job specifications for students in Psychology]. Paper presented at the Bühler-Kolloquium of the TU-Dresden. Retrieved May, 10th, 2007, from http://www.psychologie.tu-dresden.de/buehler/ew.pdf

Wilhelm, O. & Engle. R. W. (Eds.) (2005). *Handbook of understanding and measuring intelligence.* London: Sage.

Wilhelm, O., Formazin, M., Böhme, K., Kunina, O., Jonkmann, K., & Köller, O. (2006). Auswahltests für Psychologiestudierende: Befundlage und neue Ergebnisse [Selection tests for Psychology students: General evidence and new results]. *Report Psychologie, 31,* 338-349.