Once Upon a Time in the Test: Sex Differences in the Prediction of Academic Achievement and Job Performance

Dissertation submitted for the degree of

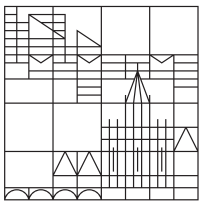
Doctor of Natural Sciences

Presented by

Johannes Schult

at the

Universität Konstanz



Faculty of Sciences

Department of Psychology

Date of the oral examination: October 21, 2013

First supervisor: Prof. Dr. Benedikt Hell

Second supervisor: Prof. Dr. Britta Renner

Contents

List of Figures

List of Tables

Acknowledgements

Thanks to my advisor Prof. Dr. Benedikt Hell and to the commitee members Prof. Dr. Britta Renner and Prof. Dr. Thomas Götz. Thanks to my co-workers in the Genderfairness project Franziska Fischer, Katja Päßler, Eunike Wetzel, and Michael Dantlgraber. Thanks to our Hiwis Lea Ludwig, Sabrina Strohmeier, Alice Stockmann, Amelie Werner, and Julia Maxie Zelfel. Thanks to the University of Konstanz for being an excellent research facility filled with many a great mind.

Thanks to Sebastian Schult for proofreading. Thanks to Christoffer Wittmann and Jan Böhnke for critical thinking in and beyond academia. Thanks to Susanne Lehner for helpful comments. Thanks to Thomas Hartman, Julian Keil, and also the Holzmannjungs for rock and roll. Thanks to Kai Müller-Berner for teaching me early on: "Intelligenz ist messbar." Thanks to Prof. Dr. Jörn Sparfeldt for keeping me busy these days.

1 Conducted studies and own research contribution

The studies of the present thesis were co-authored and supported by a number of colleagues, who are listed below along with my own research contributions.

1.1 Study 1: Sex-Specific Differential Prediction of Academic Achievement by German Ability Tests

Authors: Johannes Schult, Benedikt Hell, Katja Päßler, and Heinz Schuler

Published in the International Journal of Selection and Assessment (Schult, Hell, Päßler, & Schuler, 2013)

I developed the research strategy jointly with Benedikt Hell. I planned and performed the statistical analyses. I drafted the manuscript. Benedikt Hell collected the data of Sample 1; Sabrina Trapmann and Benedikt Hell collected the data of Sample 2; Katja Päßler and Benedikt Hell collected the data of Sample 3; Heinz Schuler supervised the data collection.

1.2 Study 2: Women and Men Tend to Use Different Narrow Abilities in Tests of Scholastic

Authors: Johannes Schult, Franziska T. Fischer, and Benedikt Hell

Submitted for publication in the Journal of Educational Measurement

I developed the research strategy. I planned the study jointly with Franziska Fischer and Benedikt Hell. I collected the data jointly with Franziska Fischer. I planned and performed the statistical analyses. I drafted the manuscript aided by Franziska Fischer and Benedikt Hell.

1.3 Study 3: Prädiktoren des Berufserfolgs von Hochschulabsolventen: Befunde aus dem Sozio-Ökonomischen Panel

Author: Johannes Schult

Published in Wirtschaftspsychologie (Schult, 2012)

I developed the research strategy. I planned and performed the analyses. I drafted the manuscript.

1.4 Study 4: Sex-Specific Differential Prediction of College Admission Tests: A Meta-Analysis

Authors: Franziska T. Fischer, Johannes Schult, and Benedikt Hell

Published in the Journal of Educational Psychology (F. T. Fischer, Schult, & Hell, 2013)

Franziska Fischer and I jointly developed the research strategy and coded the primary studies. I led the planning of the statistical analyses. Benedikt Hell conceived and supervised the study; Franziska Fischer ran the analyses and drafted the manuscript aided by Benedikt Hell and me.

2 General Introduction

High-stakes tests loom large. Millions of young people's lives are affected by college admission test scores every year. In order to select their future students from the pool of applicants, institutions of higher education rely on indicators that can predict subsequent academic performance. Two such indicators are commonly used: previous scholastic achievement and scholastic aptitude. The former is often summarized in a grade point average (GPA) whereas the latter is assessed by specifically designed tests.

Meta-analyses have repeatedly shown the predictive validity of high school GPA (HSGPA; Bejar & Blew, 1981; Schuler, Funke, & Baron-Boldt, 1990; Trapmann, Hell, Weigand, & Schuler, 2007; Richardson, Abraham, & Bond, 2012) and scholastic aptitude test scores (Bejar & Blew, 1981; Kuncel, Hezlett, & Ones, 2001; Donnon, Paolucci, & Violato, 2007; Hell, Trapmann, & Schuler, 2007; Kuncel, Credé, & Thomas, 2007; Richardson et al., 2012). Effect sizes tend to be medium-sized and may be augmented by correcting for restriction of range and measurement error (Oh, Schmidt, Shaffer, & Le, 2008; Sackett & Yang, 2000).

Before considering the main topic of this thesis, gender fairness, it should be helpful to introduce in more detail the three major variables used in these studies: college admission test scores, HSGPA, and academic achievement.

College Admission Testing China is currently the country with the most test takers, but research regarding the Chinese National College Entrance Exam (NCEE) is scarce (Bai & Chi, 2011). Most studies of admission tests originate from the United States, a country with a long tradition of college entrance exams (Atkinson & Geiser, 2009; Zwick, 2002). There is a multitude of national tests that range from general admission exams for first-time students (e. g., SAT, ACT[1]) to subject-specific graduate exams (e. g., GRE[2],

[1]Both, SAT and ACT are no longer abbreviations. Instead, they comprise a broad set of tests. Aptitude tests aimed at high school graduates are prevalent, but there are also other test services offered, e. g., subject-specific knowledge tests.

[2]Graduate Record Examinations

GMAT[3], MCAT[4], LSAT[5]). They usually feature multiple subtests covering different aspects deemed necessary for academic success. Some of these subtests aim directly at specific content areas whereas others assess more general constructs. The tests are often developed and marketed by nonprofit organizations in order to prevent conflicts of interest. Still, test transparency is limited because of the high-stakes nature of the tests and the need to repeat items in order to ensure that tests are equally difficult and thus comparable (ACT, 2012).

There is another measurement issue beside the constant change and update of test content. What are college admission tests supposed to measure? Scholastic aptitude is a poorly defined construct (F. Patterson & Ferguson, 2010). It includes all cognitive and noncognitive aspects that are deemed relevant for successfully graduating from college. Admission tests do not attempt to cover this possibly infinite set of constructs. Indicators that are susceptible to faking are usually excluded along with indicators that offer only little incremental validity. Instead, test makers aim at assessing analytical writing and reasoning ability in an academic context (Frey & Detterman, 2004). Unsurprisingly, there is a substantial correlation between IQ and admission test scores (K. A. Koenig, Frey, & Detterman, 2008). Test scores have even been used as measures of general intelligence in some instances (e. g., Jackson & Rushton, 2006). All general college admission tests share this property. Still, it is important to not regard tests as interchangeable, even though different test scores tend to correlate highly (Frey & Detterman, 2004).

School Grades The correlation between IQ and school performance is about .5 (Neisser et al., 1996). This highlights the role of intelligence in educational success. Still, school performance and academic performance, as captured by grades, defy a clear measurement concept. The long stretch of time across which grades are gathered makes up for the lack of explicit

[3]Graduate Management Admission Test
[4]Medical College Admission Test
[5]Law School Admission Test

psychometric theory. As long as each grade measures cognitive ability at least partially, GPA becomes a more reliable indicator of cognitive ability with each additional measurement.

There are additional factors that play a role in determining scholastic achievement, contributing to the high validity of HSGPA in the prediction of college performance. Conscientiousness is the most promising candidates among noncognitive predictors of academic performance (Poropat, 2009; Trapmann, Hell, Hirn, & Schuler, 2007). HSGPA is thus a composite of intelligence and behavior which is relevant academically (Allen, Robbins, Casillas, & Oh, 2008).

A way to deal with the lack of a psychometric foundation is modeling latent ability factors based on manifest grades (Deary, Strand, Smith, & Fernandes, 2007). This is satisfying from a measurement point of view, but somewhat out of touch with reality, because college degrees do not come with scores on latent ability scales. It is still the actual grades (or sometimes ranks) that count, which is why manifest GPAs are commonly used in research. Standardizing grades within institutions or regional units helps countering issues like grade inflation (Bejar & Blew, 1981) and grading styles (Bridgeman, McCamley-Jenkins, & Ervin, 2000).

Academic Success Academic success has many faces (Trapmann, 2008). Degree completion indicates that a person has passed all necessary exams. This binary outcome measure is related to retention, i. e., the continuation of university studies at a given point in time. Retention is easier to assess, because one does not have to wait for students to finish their studies. Instead, a student cohort can be probed when a certain amount of time has passed since they took up their studies.

The most prevalent success criterion is, however, college grades. Cumulative college GPAs are preferred as they contain the most information, but due to time constraints first-year GPA (FYGPA) is used as a surrogate outcome in the majority of studies (see Study 4 in Section 6). FYGPA is, in turn, predictive of cumulative CGPA at the end of university (Allen et al., 2008; Sackett, Borneman, & Connelly, 2008).

There are further aspects of academic success like time to graduation, satisfaction, and University Citizenship Behavior, but those are less frequently used in studies of college admission testing (Trapmann, 2008). The sole focus on objective outcomes is certainly a limited one that excludes other benefits of college attendance (Stemler, 2012). For example, personal maturation is difficult to assess but might be important for a person as a whole. Also, future job performance is not just a matter of grades and degrees; social skills and metacognitive abilities may also help.

2.1 Test Fairness

Test Fairness is a broadly used term that requires a clear definition if it is to be used in research. The topic remains heterogeneous even if only the fairness of psychological tests is considered. The *Standards for educational and psychological testing* (American Educational Research Association [AERA], American Psychological Association [APA] and National Council on Measurement in Education [NCME], 1999) have been repeatedly updated and provide a multitude of guidelines. They suggest four major aspects of fair test use:

- equal test scores across subgroups,
- equal opportunities to learn,
- equal treatment across subgroups (e. g., testing conditions, practice material, feedback), and
- lack of bias.

The first and the last point are related to psychometric test properties, whereas the others pertain to procedural aspects. The present thesis deals with the psychometric aspects of test fairness. Here, the lack of bias is the key feature of test fairness.

The notion of equal test scores across subgroups as a prerequisite for test fairness is disputed (AERA, APA, & NCME, 1999). In practice, it corresponds to a system of group-specific quotas, which can be at odds with

other aspects of fairness. If there are subgroups in a pool of applicants that do differ on a variable of interest, the most valid selection procedure is likely to mirror this skewed distribution (Meade & Fetzer, 2009). From a psychometric point of view, equal test scores across subgroups are not necessary for a testing procedure to be fair (AERA, APA, & NCME, 1999).

The main definition of test fairness in the present thesis corresponds to the absence of bias (Cleary, 1968; Meade & Tonidandel, 2010). It is important to note that test scores cannot be biased per se. Only in a particular context can they be regarded as biased (Darlington, 1971; Meade & Fetzer, 2009). The term test bias can therefore be misleading, because it suggests an all-encompassing problem with a given test, although it usually relates to specific test properties in a particular setting.

Bias in admission testing maifests itself often in one (or more) of the following indicators (AERA, APA, & NCME, 1999):

- differential item functioning (DIF),
- differential validity, and
- differential prediction.

The statistical terms for their absence are "measurement invariance" (no DIF) and "predictive invariance", respectively (Millsap, 2007). In practice, test fairness is not a dichotomy, although significance tests with the null hypothesis that the group effect is zero help maintain this illusion. The question is not really whether there is a group difference or not[6]; the question is how large the bias is in a particular admission setting, and which settings show a comparable amount of bias.

Therefore, the following description of the three bias manifestations listed above focuses on the extent of bias (e. g., in terms of effect sizes), although significance testing is not completely ignored.

[6]As Cohen (1990, p. 1308) put it: "The null hypothesis, taken literally (and that's the only way you can take it in formal hypothesis testing), is *always* false in the real world.".

2.1.1 Differential Item Functioning

DIF occurs when the item response is not only a function of a person's (latent) ability but also depends on group membership or other factors (de Ayala, 2009, pp. 323–345). Item Response Theory (IRT) provides a clear basis for the concept of DIF. There are also other, usually less rigid and more robust approaches beyond this framework.

The Rasch model is one of the most parsimonious IRT models, estimating only one parameter per item (i. e., the item difficulty). Here, DIF means that an item has different difficulty parameters for subgroups. IRT models with two parameters add an individual item discrimination parameter for each model. They can be used to also identify non-uniform DIF (i. e., group differences in an item's discrimination parameter).

The presence of DIF suggests that an item is unfair, but it provides no definite conclusion. One should rather evaluate the content of each DIF-item in order to identify potential causes—and thus bias.

Further complications arise because several different methods for detecting DIF have been developed—some within the IRT framework (e. g., differences in difficulty parameters), others based on simpler models (e. g., Mantel-Haenszel)—and they are not always in agreement (Abedalaziz, 2010). Mechanical exclusion of DIF-items (purification) is a way to reduce bias, but keeping DIF-items in a long test and balancing DIF across groups is sometimes preferred over item deletion and the associated loss of precision (Osterlind & Everson, 2009).

In college admission testing, items that show DIF are usually identified in pretests and then revised or discarded in order to obtain a measurement that is invariant across subgroups like sex and ethnicity (Curley & Schmitt, 1993; Zhang, Dorans, & Matthews-López, 2005). Therefore, the actual test items tend to exhibit minimal DIF (Lawrence, Curley, & McHale, 1988). DIF analysis of the scholastic aptitude tests used in Study 2 (Bundesagentur für Arbeit, 2004a, 2004b) show a similar pattern. The amount of items that show sex-related DIF is below that expected by chance.

A balanced set of items is usually just the starting point for further test

fairness investigations. Invariance regarding criterion validity is the next step.

2.1.2 Differential Validity

Indicators used in the college admission process must show substantial criterion validity. In addition to this basic property, similar validity coefficients for subgroups are desirable in order to base the admission decision on a common prediction model (Holden, 1989; Kuncel et al., 2007).

To test for differential validity, correlations are Fisher z transformed (Bortz & Döring, 2006, p. 611),

$$Z = \frac{1}{2} \ln |\frac{1+r}{1-r}|. \tag{2.1}$$

Then the difference $q = Z_f - Z_m$ is the test statistic, which is normally distributed with standard error

$$\sigma_q = \sqrt{\frac{1}{n_f - 3} + \frac{1}{n_m - 3}}, \tag{2.2}$$

where n_f and n_m are the sample sizes for women and men, respectively. The 95 % confidence interval is $q \pm 1.96 \cdot \sigma_q$ (Trattner & O'Leary, 1980; Weaver & Wuensch, 2013).

There appears to be a small but consistent amount of differential validity in U. S. American samples for HSGPA, admission test scores, and composites of HSGPA and test scores (Mattern, Patterson, Shaw, Kobrin, & Barbuti, 2008; Kuncel et al., 2007; Young & Kobrin, 2001; Bridgeman et al., 2000). Overall, differential validity tends to be lowest for predictions based on HSGPA only (.02), and largest for predictions based on admission test scores (.08; Mattern et al., 2008). For some college majors (e. g., journalism), validity coefficients differ up to .20 between men and women (Shaw, Kobrin, Patterson, & Mattern, 2012). Testing for differences as small as these requires large sample sizes (Trattner & O'Leary, 1980). Therefore, studies with medium sample sizes like Study 2 tend to find similar effects, but lack the statistical power for consistent significant results.

Differential validity is not entirely independent from differential prediction, because regression slopes are related to the validity coefficients. Still, a clear conclusion from differential validity to differential prediction is rarely possible under realistic conditions (Millsap, 1995, 2007).

2.1.3 Differential Prediction

Cleary (1968, p. 115) provided the seminal definition of differential prediction, which is still widely used (see Aguinis & Smith, 2007; Meade & Tonidandel, 2010):

> A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup.

Thus, predictor scores are Cleary-fair when valid predictions for all subgroups can be made using a common regression line. Differential prediction denotes group-specific under-prediction or over-prediction. Statistical test bias in this narrow sense jeopardizes the integrity of aptitude tests and other admission criteria.

Previous studies suggest that standardized tests tend to underestimate the academic achievement of women (Mattern & Patterson, 2013; Young & Kobrin, 2001). This sex-specific differential prediction can be explained at least partially by course-taking patterns (Ceci, Williams, & Barnett, 2009; Sackett et al., 2008).

An overview of statistical approaches to differential prediction is given in the next section.

2.2 Excursus: Statistical Approaches to Differential Prediction

The analysis of differential prediction through moderated multiple regression (MMR) is outlined in the next section along with a discussion of its merits and pitfalls. After that, the features of regression residuals are described. They may serve as an addition or in some cases—for example, in meta-analysis—even as an alternative to MMR.

2.2.1 Moderated Multiple Regression

The *Standards for educational and psychological testing* (AERA, APA, & NCME, 1999, p. 82) recommend that empirical studies of differential prediction "should include regression equations (or an appropriate equivalent) computed separately for each group or treatment under consideration or an analysis in which the group variables are entered as moderator variables." This is in line with the method originally employed by Cleary (1968), which led to MMR being labeled the "Cleary model". It is noteworthy that comparing separate regression lines and using moderator variables are equivalent procedures (Bartlett, Bobko, Mosier, & Hannan, 1978). The former approach pays tribute to the psychological tradition (Gulliksen & Wilks, 1950) whereas the latter reflects the progress in statistical computing witnessed, for example, in econometrics, where the test for differential prediction is also known as Chow test (Dougherty, 2007). The formula for transforming the regression equations is (Wooldridge, 2006)

$$\left.\begin{array}{l} \hat{Y}_f = b_{0f} + b_{1f}X_{1f} \\ \hat{Y}_m = b_{0m} + b_{1m}X_{1m} \end{array}\right\} \quad \hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_1X_2 \tag{2.3}$$

with $b_0 = b_{0m}$, $b_1 = b_{1m}$, $b_2 = b_{0f} - b_{0m}$, $b_3 = b_{1f} - b_{1m}$ and X_2 as dummy variable for group (e. g., sex). X_1 is the predictor (e. g., admission test score) and $\hat{Y}$ is the predicted criterion (e. g., predicted academic achievement). The t-test for the coefficient of the interaction term, b_3, corresponds to the tests of equal slopes; the t-test for b_2 corresponds to the subsequent test of equal intercepts (Nagl, 1992).

The last decade saw MMR with a combined regression equation emerge as the preferred way of analyzing differential prediction compared to separate regression equations for each subgroup (Schult, Fischer, & Hell, 2010).

MMR in Practice In practice, the analysis of differential prediction requires three prior steps in order to identify potential causes (and later on remedies) for predictive bias (Meade & Tonidandel, 2010): (1) examination of differential item functioning (DIF) and differential test functioning, (2) examination of significant group mean differences on test and criterion scores, and (3) evaluation of d effect size estimates for these differences.

The actual MMR analysis begins with the test for slope differences. It continues with the test for intercepts differences in case there are no significant slope differences. The regression lines are assumed to be identical and the test instrument is considered Cleary-fair if none of these tests achieve significance. The test results alone (i. e., p-values) do not provide sufficient information about the direction of the bias. Even the parameter estimates do not always yield a clear picture when slope differences are present. Another prerequisite has been proposed by Gulliksen and Wilks (1950). Standard errors of estimates should be equal across groups. This hypothesis should be tested before probing slope and intercept differences. The additional test has been performed in some studies (e. g., Thomas, 1979) but not in other, more recent ones (e. g., Bridgeman & Wendler, 1991). Ambiguities like this lead to varying degrees of clarity and detail when differential prediction is reported, impeding the aggregation of results across studies.

Scaling Issues in MMR In the full MMR model, the main effect of the dummy variable depends on the scaling of the predictor variables (Schmidt & Hunter, 1982), because it is assessed at the point where the predictor is 0. For example, SAT scores range from 600 to 2400; here the intercept test would refer to an impossible test score. The recommended solution is to center all continuous variables but not the dummy. One can also identify the test score range that is Cleary-unfair, i. e., in which the main effect is significant (Aiken & West, 1991). Still, a literature search on sex bias in

college admission tests[7] yielded only one study that reported these ranges of significance (Patton, 1998). Unfortunately, the predictors in most empirical studies of differential prediction are not centered, which may lead to false conclusions as even the sign of the bias can change depending on the scaling of the independent variable.

MMR Effect Size and Test Power There are several MMR effect sizes. Most prominent are η^2, the proportion of variance explained by a variable, partial η^2, the proportion of variance that is explained exclusively by a variable but not by other variables in the model, and f^2, the ratio of variance explained by the moderator to unexplained variance in the criterion (Cohen, 1988; Aiken & West, 1991, p. 157). η^2 can be calculated from f^2. See Cohen (1988, pp. 281–283) for a conversion table. A comprehensive review of MMR findings in psychology showed a median observed effect size $f^2 = 0.002$ (Aguinis, Beaty, Boik, & Pierce, 2005), which is much lower than what Cohen (1988) denotes as small ($f^2 = 0.02$). As a consequence, the detection of moderator effects in small samples (< 100) will rarely be fruitful. Sample size should be adjusted (i. e., increased) for models with multiple predictors[8]. Complex models are unsuitable for small studies with an insufficient number of participants. Test power can be further diminished by issues like unequal subgroup sizes and scale coarseness (Aguinis, 2004). Unfortunately, even very small effect sizes usually reflect real and relevant issues in academic and economic placement decisions.

Conclusion MMR is the prevalent method for investigating predictive bias. The method is available in all general software packages. MMR is a common approach and certainly a helpful way to analyze differential prediction. Still, it has its limitations due to issues like test power, subgroup size, and model complexity.

[7] See Section 6.2.1 in Study 4 for details of the literature search.

[8] Sample size calculations can be performed using the program MMRPOWER located at `http://mypage.iu.edu/~haguinis/mmr/mmrpower/MMRPower.html` (Aguinis, Boik, & Pierce, 2001).

2.2.2 Group-Specific Residuals

The shortcomings of MMR led to a search for more robust and more practical ways to detect predictive bias. Possibly the most widespread alternative to MMR is the analysis of group-specific residuals. The residuals stem from a common regression line that is fit to the whole sample ($\hat{Y} = b_0 + b_1 X_1$). Next, the mean residual $\bar{E}_j = \bar{Y}_j - \bar{\hat{Y}}_j$ is computed for each group $1, \ldots, j$. Nonzero mean residuals indicate the presence of test bias, according to the definition of Cleary (1968) quoted at the beginning of this article. Positive errors denote underprediction whereas negative errors indicate overprediction.

The College Board, best known for administrating the SAT, utilizes mean residuals in most of its recent reports on test fairness (e. g., B. F. Patterson, Mattern, & Kobrin, 2009; Mattern et al., 2008). Given the large sample sizes ($n > 150,000$), significance tests are no longer a viable option. Reporting residuals helps communicating the test properties to a lay audience without abandoning statistics all together. Unstandardized mean residuals can be easily interpreted as the average deviation from the common prediction in the unit of the criterion scale.

Related studies with smaller sample sizes report mean residuals, as well (e. g., Sireci & Talento-Miller, 2006), although no common procedure has been established until now. The t-test we present here is one more step from there. It provides attractive features, including a way to aggregate findings in meta-analysis.

Testing for Nonzero Mean Residuals The null hypothesis suggests that the test in question is fair, i. e., $H_0 : \bar{E}_j = 0$. A simple t-test for a two-group scenario

$$T = \frac{(\bar{E}_1 - \bar{E}_2)}{S_E} \cdot \sqrt{N} \tag{2.4}$$

was proposed by Lawshe (1983). Unfortunately, the two mean residuals are not independent from each other, thus violating an assumption of the test. This can be rectified by a using a t-test for group-specific errors,

$$T = \frac{(\bar{E}_j - 0)}{S_{E_j}} \cdot \sqrt{N_j}, \tag{2.5}$$

where N_j is the sample size of subgroup j. It tests the null hypothesis that the deviation of group j's mean performance does not differ from the value predicted by a common regression line. It can be argued that the overall deviation of residuals S_E should be used instead of S_{E_j}. This helps reducing the measurement error if the assumption holds that both groups' residuals come from the same distribution.

The test should not be performed independently for every group because using the ordinary least square estimator yields a total mean residual of zero and accordingly

$$\sum_{1}^{j} N_j \bar{E}_j = 0. \tag{2.6}$$

In other words, if one half of the sample has a positive mean error the other half must have a negative mean error[9].

As a consequence of Equation (2.6), $\bar{E}_j$ is not entirely independent of subgroup size. This statistical flaw mirrors the existing predictive bias, which—on average—affects smaller subgroups more than large ones. The impact (be it positive or negative) of differential prediction on minorities is larger than the impact on the majority. Low test validities are another source of possible distortion. They inflate the standard deviation of the residuals, rendering the t-test insensitive to small predictive bias.

Mean residuals indicate the average unfairness across the whole test score range per definition. Differential prediction on subsections of the test score range does not necessarily reflect in the mean residuals if group-specific regression lines intersect near the sample's centroid (Norborg, 1984). To remedy this problem one can either resort to the MMR procedure or—which should be done in any case—inspect the scatter plots containing group-specific regression lines to determine the potential impact on the region of interest, e. g., around a cut-off point used for the admission of students.

Conclusion Analyzing residuals provides easy-to-use conclusions regarding test bias, especially if complex regression equations with multiple predic-

[9]It is possible to test the numerator of Equation (2.4), $\bar{E}_1 - \bar{E}_2$, using a bootstrap approach, but this renders the method difficult to use for lay persons.

tors are used. Depending on the context, either the unstandardized mean residual or the associated effect size illustrate the overall test bias in a way that requires little knowledge of statistics. This can be crucial when people with limited assessment literacy like school personnel interpret and utilize the results (Zwick et al., 2008).

2.2.3 Reconciling MMR and Residuals

Estimating group-specific parameters with MMR highlights the between-group comparison whereas the analysis of group-specific residuals from a common regression line emphasizes the extent of over- and underprediction in a way that is more readily available to nonstatistical users.

MMR remains the method of choice. Still, reporting mean residuals—preferably along with their standard deviation S_{E_j}—facilitates the aggregation of findings from different sources (see also Section 2.3.4). The analysis of residuals can be useful with large sample sizes[10] and multiple predictors.

2.2.4 Visual Inspection

Both approaches usually benefit from graphical data analyses that augment numerical statistics. Violations of assumptions of linearity can often be spotted in regression diagnostic plots (Schnell, 1994). Scatterplots with fitted values and residuals on the respective axes are not only useful to check assumptions regarding the distribution of residuals (Hamilton, 1992); they also facilitate the evaluation of the predictive validity in test score ranges of interest.

Traditional MMR analysis should always be accompanied by appropriate graphics. Two-way scatterplots featuring regression lines are an easy way to visualize differential prediction. Lines of nonlinear fit and residual plots are helpful to detect weaknesses of the MMR model. These graphics should be inspected regardless of what type of analysis is performed on the data. The

[10]Current reports on large scale validation studies usually contain mean residuals but not MMR (e. g., B. F. Patterson et al., 2009).

study by J. A. Koenig, Sireci, and Wiley (1998) is a good example of how supplementary plots enhance the understanding of the data.

Visual inspection is also helpful for checking the underlying assumption of linearity, which is necessary for all models presented here. Influential outliers, especially in small subgroups, can be identified by inspecting scatterplots and also by looking at measures like Cook's distance (Fox, 1991). If basic regression assumptions do not hold, subsequent investigations of differential prediction—at least for the whole sample—are no longer tenable.

2.3 Open Questions Regarding Gender Fairness

A large body of research about the fairness of college admission procedures has been accumulated over the past decades. Major questions regarding sex-specific DIF, differential validity, and differential prediction have been studied with considerable success. Some insights led to significant changes (e. g., the introduction of essay writing in the SAT to augment the mathematical and verbal reasoning subtests); others indicated that the status quo of a particular test provides an acceptable degree of fairness.

Despite this plethora of findings, there are some areas where no consensus has been reached yet and further study is needed. Also, college admission research is an ongoing challenge as a new generation of applicants come to university each year to receive an academic education and, in the end, a degree. Since 1988, women have outnumbered men in U. S. American colleges; about 57 % of all currently enrolled students are women (Snyder & Dillow, 2012). In Germany, more students are male (about 53 %), but there are slightly more women who graduate successfully (51 %; Statistisches Bundesamt, 2012). The fair distribution of available places is an important issue in this context.

The first open question addressed in this thesis is whether the test fairness of cognitive tests that are administered to German students mirrors U. S. American findings.

2.3.1 Test Fairness in Germany

Tests of scholastic aptitude play an important role in the U. S., where competition is fierce among elite universities and the heterogeneity of school districts diminishes the validity of high school grades. German universities traditionally rely on HSGPA. Still, subject-specific admission tests may be used to select students (Heine, Briedis, Didi, Haase, & Trost, 2006). Medicine is the only subject that has seen the repeated nation-wide use of aptitude tests in the admission process (Trost, Nauels, & Klieme, 1998; Kadmon, Kirchner, Duelli, Resch, & Kadmon, 2012). In recent years, the Bologna Reform led to a decentralization of student placement. Faced with the challenge to make the admission decisions themselves, some universities employ standardized tests (along with other tools), mainly in fields of study where there are more applicants than available places (e. g., Formazin, Schroeders, Köller, Wilhelm, & Westmeyer, 2011). Other institutions, for example private colleges (Dlugosch, 2005), also use tests in their admission procedure.

To see how well findings from the USA can be translated to Germany, data from three student samples were analyzed in Study 1 (see Section 3). The relative lack of general admission tests in Germany makes it difficult to assess external validity. Furthermore, there are almost no native tests of scholastic aptitude in use that are not subject-specific[11]. In order to study test fairness in Germany one has three options:

- translate foreign test items,
- develop a new test, or
- use tests that are closely related to admission tests.

Translating existing tests is difficult, because items are rarely published and are subject to strict copyright restrictions. Developing a new test requires substantial effort in terms of money, manpower, and time. This leaves the third option: use similar tests.

[11]Notable exceptions are the Test der akademischen Befähigung (TAB) and the Auswahltest der Studienstiftung (ATS; cf. Trost, 2003).

Existing tests that are subject-specific may show conceptual overlap with general tests. Still, they are tailored to assess abilities, and potentially skills or knowledge, that pertain to constructs that are typical for the field of study at hand. It is possible that subject-specific tests are sufficiently valid if they were used in a general setting. However, face validity would be low in some instances, for example if a law school test is presented to prospective physics students. Fortunately, another set of tests that has plenty in common with scholastic aptitude tests is readily available: tests of general mental ability—or in short: intelligence tests.

There is a large conceptual overlap between scholastic aptitude tests and intelligence tests (Frey & Detterman, 2004; K. A. Koenig et al., 2008) and SAT scores have even been used as indicators for general mental ability (Jackson & Rushton, 2006). In return, intelligence tests have been successfully employed in student admission (Sternberg, Bonney, Gabora, & Merrifield, 2012). The main link between tests is reasoning (Zwick, 2007).

Intelligence test scores and scholastic aptitude test scores are not equivalent—SAT scores predict college grades beyond IQ (Coyle & Pillow, 2008)—and German laws actually prohibit the use of tests of general intelligence in college admission. Despite these flaws, tests of general mental ability remain a valuable tool to investigate the differential prediction of German students' academic achievement, in particular because their construct validity has been studied thoroughly.

The findings of Study 1 suggest that differential prediction is related to facets of reasoning (see Sections 3.4 and 7 for additional discussion). To investigate the role of intelligence facets in more detail, in particular in combination with a test of scholastic aptitude, longitudinal data was gathered for Study 2, which is introduced in the next section.

2.3.2 Construct Validity and Criterion Validity

The structure of college admission tests is similar to models of human intelligence where more general constructs appear side by side with more detailed facets. In most intelligence tests, it is possible to extract a general factor from

subtest scores that is commonly labeled g (Jensen, 1998). This overarching factor of general mental ability goes back to the early days of intelligence research (Spearman, 1904). The idea of g can be regarded as either pleasantly parsimonious or overly simplistic, because it covers some of the test score variance but not all and it reduces people's mental ability to a single number on the IQ scale. Hierarchical models with correlated group factors underneath g have been developed in order to differentiate intraindividual strengths and weaknesses (Carroll, 1993; Neisser et al., 1996). The Cattell-Horn-Carroll (CHC) theory of intelligence is a synthesis of two major factor models (McGrew, 2009). It retains Carroll's (1993) fluid and crystallized intelligence (Gf–Gc) and provides a taxonomy of further broad abilities like short-term memory, visual processing, processing speed, reading and writing, and quantitative knowledge (McGrew, 2009).

Unlike intelligence, scholastic aptitude is a blurry concept. It is supposed to comprise abilities and skills necessary for academic success (College Entrance Examination Board, 2004). This notion already implies the link to future performance. Consequently, test makers aim at high criterion validities first and foremost. Some college admission tests strive to assess actual achievement rather than intellectual ability in order to boost criterion validity (Zwick, 2007).

Academic success defies a clear operationalization; studies tend to use proxies like first-year GPA (Stemler, 2012). This contributes to the confusion surrounding the question "What do college admission tests measure?" (Zwick, 2007, p. 11).

The first goal of Study 2 is to assess how facets of reasoning relate to scholastic aptitude. Both constructs are modeled as latent factor variables in order to reduce measurement error. A strong relationship between intelligence and scholastic aptitude is to be expected (Frey & Detterman, 2004). In terms of subfactors (i. e., verbal, numeric, and figural content), the data will shed light on a part of college admission testing that is yet to be explored.

The structural equation models presented by Coyle and Pillow (2008) suggest that the predictive validity of the SAT and the ACT rely on more than just g. But is cognitive ability sufficiently reflected in admission test

scores (Zwick, 2007)? Study 2 provides an empirical answer to this question by testing whether the predictive validity of intelligence facets is fully mediated by scholastic aptitude test performance.

The second goal of Study 2 is to explore sex differences in both construct validity and criterion validity. Previous research suggests higher criterion validity for women (Young & Kobrin, 2001). Sex differences in some intelligence facets (Nisbett et al., 2012; Ellis et al., 2008) might be reflected in the construct validity of the scholastic aptitude tests.

Furthermore, Study 2 complements Study 1 by using subject-specific tests of scholastic aptitude instead of intelligence tests. First-year GPA is used as a criterion for academic success. More distal outcomes are considered in the next study.

2.3.3 A Look Beyond College

Previous scholastic achievement and noncognitive factors can be used to predict job performance (Roth, BeVier, Switzer III, & Schippmann, 1996; Roth & Clarke, 1998; Judge, Higgins, Thoresen, & Barrick, 1999; Judge, Heller, & Mount, 2002). The underlying prediction models are basically the same as the ones used in validation studies of scholastic aptitude tests. So what happens when we move beyond college graduation and look at the differential prediction of job performance?

First of all, there are even more predictors and outcome variables than in college admission testing. A look at selection instruments used by human resources departments confirms this notion: there are various kinds of interviews, application documents, assessment centers, references, achievement tests, intelligence tests, personality tests, work samples, and medical opinions—to name but the ones that are most frequently used[12] (Schuler, Hell, Trapmann, Schaar, & Boramir, 2007). General mental ability is among the most valid predictors of job performances and permeates most of the instruments listed above (Kuncel, Wee, Serafin, & Hezlett, 2010; Schmidt &

[12]Internal selection and placement decisions are most often based on the judgment of supervisors and interviews (Hell, Boramir, Schaar, & Schuler, 2006).

Hunter, 1998). There are also noncognitive predictors like interests and personality that can play a crucial role in career attainment (Chapman, Uggerslev, Carroll, Piasentin, & Jones, 2005; Judge et al., 2002). Each predictor's ability depends on the choice of criteria.

Career success can be divided into extrinsic and intrinsic factors (Judge et al., 1999). Extrinsic factors are aspects that can be readily observed whereas intrinsic factors like job satisfaction and recognition from others relate to a person's subjective work experience (Judge, Cable, Boudreau, & Bretz, Jr., 1995). Previous achievements are valid predictors of extrinsic career success (Judge et al., 2002) whereas motivational factors are good predictors for job satisfaction (Judge et al., 1999).

In Study 3, mental ability (using grades as proxy) and personality traits are used to predict two types of work success criteria: income (extrinsic) and job satisfaction (intrinsic). This decision was driven partly by the need to restrict the number of outcome measures to a number that is manageable and partly by the available data, which contains only a limited set of psychological measures. The data in question come from the German Socio-Economic Panel (SOEP), which has been running since 1984 and contains over 22,000 individuals at the moment (Schupp, 2009). The longitudinal design offers a compelling way to study predictive validities.

Studies of the career attainment of university graduates are usually based on cohorts of a particular institution and thus limited to the specific conditions of that institution (e. g., Abele & Spurk, 2009). Another problem with longitudinal studies is attrition. While it is practically impossible to prevent some people from dropping out of the study, auxiliary information can be used to adjust the weights of observed cases (Kalton & Flores-Cervantes, 2003). The SOEP has the advantage that it is based on random household samples from the German population. Information from previous waves can be used to estimate retention probabilities, which in turn can be inversed and serve as weights (Kalton, 1986; Kroh, 2010). The nation-wide survey structure makes it easier to track participants and conduct personal interviews after the transition from university to the labor market.

University degrees are associated with higher status and increased salaries

(Gebel & Pfeiffer, 2010). Previous research suggests that cognitive ability has predictive power beyond educational attainment (S. Anger & Heineck, 2010). Are there further factors that may be used in selecting job applicants from a pool of university graduates? And what factors may play a role when graduates choose a particular job career over another? In order to answer these questions, the validities of several noncognitive predictors along with grades are explored in Study 3. Outcomes are money and job satisfaction two years after graduation. Entering sex into the prediction models raises another question: Does differential prediction persist beyond graduation?

At this point it is important to reiterate that differential prediction does not necessary indicate a problem with the predictor. Differential prediction is tied to a particular use of selection tools and may as well indicate issues with the criterion. With regard to income, differential prediction basically points to the so-called gender pay gap—women earning less money than men with equal qualifications (C. Anger & Schmidt, 2010).

2.3.4 The Aggregation of Test Fairness Studies

The final study in this thesis returns to the sex-specific differential prediction of academic performance, which has been thoroughly studied—which makes it a perfect candidate for a comprehensive meta-analysis.

So far, there has been only one attempt of meta-analyzing published differential prediction results in higher education (Sanber & Millman, 1987). Effect sizes were derived from the t-tests of each MMR coefficient. This is statistically feasible but yields results that are at best difficult to interpret. The main conclusion is that standardized achievement tests as predictors of academic performance are unfair. Just to what extent and to which group remains unclear, because the information regarding the relationship between intercept and slope tests within each study—which becomes aggregated into one meta-relationship—is likely to vary across studies. Add to this the distortion of intercept tests by the lack of centering, and almost no useful information is left.

Three strategies for aggregating differential prediction studies are intro-

duced and discussed in the following sections: MMR, multiple regression without interaction terms, and—possibly the most promising approach—the analysis of residuals.

Meta-Analysis of MMR Results Meta regression models which are adequate for MMR data have been developed and published recently (Bowman, 2011; Aguinis, Culpepper, & Pierce, 2010). Despite having the desired statistical properties in theory, they cannot be applied to results from primary studies yet, because they require data like covariance matrices[13] which are rarely reported in publications. Mattern and Patterson (2013) recently performed a meta-analysis of the sex-specific differential prediction of academic success, combining data from various SAT validity studies. Apart from this exception (where the authors had access to the raw data), the meta-analysis of sex-specific differential prediction of academic achievement based on MMR remains unfeasible for the time being (Borneman, 2010).

Effect Size d and Meta-Analysis Given the lack of published data required for aggregating MMR results, an alternative approach is needed. The analysis of residuals (see Section 2.2.2) provides a simple solution to this problem. The effect size for the t-statistic in Equation (2.5) is

$$d = (\bar{E}_j - 0)/S_{E_j}, \tag{2.7}$$

where 0 is the mean postulated by the null hypothesis (Cohen, 1988). The analysis of residuals is not immune to test power issues when effect sizes are small, but unlike in MMR, the number of predictors does not diminish the test power. There is only one error variable no matter how complex the prediction model.

The standard process of accumulating d values like the one obtained from the t-test is a "bare-bones" meta-analysis (Hunter & Schmidt, 2004).

[13] In order to allow for a meta MMR model, covariance matrices that include covariances between the interaction variable (X_1X_2) and all other variables (including the group variable) are required. So far there has not been a single differential prediction study published that provides this information.

Meta-Analysis of Regression Data – Regression of Meta-Analysis Data Performing a meta-analysis using a multiple regression model without the interaction term is another possible solution to aggregate existing studies of differential prediction. Zero order correlations can be conventionally aggregated in case they are known. This yields a correlation matrix which—in case the correlations prove to be homogeneous—can be used to run a "synthesized" regression (Lipsey & Wilson, 2001). The resulting regression equation is—like most meta-analysis results—standardized.

Given correlations between sex and the other variables, a joint regression line can be estimated in order to test for differential prediction. Using group-specific means and standard deviations of a variable (e. g., CGPA), one can calculate the point biserial zero order correlation between X_2 (sex) and Y (CGPA; Magnusson, 1967, pp. 198–202):

$$r_{X_2Y} = \frac{(\bar{Y}_m - \bar{Y}_f)S_{X_2}}{S_y} \tag{2.8}$$

with

$$S_{X_2} = \sqrt{p_f p_m} \tag{2.9}$$

where p_f and p_m are the proportional group sizes. For correlations, r can be directly used as effect size estimate. The estimated variance of r is (Rosenthal, 1994, p. 238)

$$(1 - r^2)^2/(n - 2). \tag{2.10}$$

After correction for artifacts, one can aggregate the zero order correlations and calculate standardized regression coefficients (Tacq, 1997, pp. 149–154):

$$\beta_1 = \beta_{YX_1.X_2} = \frac{r_{YX_1} - r_{X_1X_2}r_{YX_2}}{1 - r^2_{YX_2}}. \tag{2.11}$$

A regression with two or more main effects (the full model) can be calculated in this fashion along with a restricted regression model that contains the same predictors except for the group variable. The models can be compared by looking at the conventional F-statistic (Tacq, 1997, pp. 113–115)

$$F = \frac{(R^2_{\text{full}} - R^2{}_{\text{restricted}})/(df_{\text{restricted}} - df_{\text{full}})}{(1 - R^2_{\text{full}})/df_{\text{full}}} \tag{2.12}$$

with the degrees of freedom of the restricted model $df_{\text{restricted}} = n - 1$ and those of the full model $df_{\text{full}} = n - \Delta k - 1$; Δk = number of additional parameters in the full model and (Tabachnick & Fidell, 2007, p. 131)

$$R^2 = \sum_{i=1}^{k} r_{YX_i}\beta_i. \tag{2.13}$$

This test can be used to probe the dummy variable for sex. The difference in R^2 further shows the amount of variance explained by sex.

An abundance of suitable data has been published in differential prediction studies. The challenge is to establish homogeneous subgroups for which meta regressions can be run subsequently. The main issues with the otherwise handy method of obtaining correlations (and, eventually, a meta regression equation) is: The estimation of a meta regression equation assumes the absence of moderators. In return, a homogeneity test can only be performed for each correlation individually (Lipsey & Wilson, 2001). Another minor issue is the standardization of the dummy variable: Standardizing a dummy variables is not recommended, because its variance is a function of its frequency distribution; see Equation (2.9) (Aiken & West, 1991).

Outlook This thesis focuses on sex differences in admission testing and higher education. The aggregation of residuals is applied in the meta-analysis in Study 4, which deals with the test fairness of scholastic aptitude tests (see Section 6).

The methods for meta-analyzing differential prediction studies outlined above can be used to aggregate finding for other groups as well as for other settings. There is a large body of research regarding ethnic minorities (Young & Kobrin, 2001) that has not been meta-analyzed so far. Here, disentangling group differences and majority–minority effects is an additional challenge (Wainer, Saka, & Donogue, 1992). Other settings include, among others, the differential prediction of school children's performance (Duckworth & Seligman, 2006) and sex differences in preemployment testing (Aguinis et al., 2010).

2.4 Studies in This Dissertation

I present four studies to illuminate the extent of bias in college admission tests and to explore possible explanations and broader consequences. Study 1 features three German student samples. It provides deeper insights into the extent of sex-specific differential prediction in Germany. The role of intelligence facets in tests' criterion validity and sex bias is explored in more detail in Study 2. Here, a sample of college freshmen took an intelligence test as well as a scholastic aptitude test from their respective field of study. Study 3 goes beyond the walls of college. Its focus lies on the prediction of the job performance of college students shortly after graduation. Finally, Study 4 provides a meta-analysis of sex-specific differential prediction in college admission testing.

3 Study 1: Sex-Specific Differential Prediction of Academic Achievement by German Ability Tests

Abstract

Tests of cognitive ability play a major role in the selection of students. Still, data regarding the fairness of standardized tests in Germany is scarce. We use three samples ($n = 2,616$; 58 % women) from German universities to investigate the sex-specific differential prediction of college performance based on intelligence tests. The predictive bias we find is small and in line with US-American research. The direction of the effect depends on the cognitive ability domain investigated: Numeric test scores are prone to disadvantage women whereas verbal test scores are more likely to discriminate against men. Including high school grade point average in the prediction model can help to offset differential prediction that underestimates women's academic achievement.

Keywords: college admission, intelligence, sex differences, differential prediction

[14]This research used data of the project 'Student Selection' ('Studierendenauswahl') commissioned by the Landesstiftung Baden-Württemberg and the Stifterverband für die deutsche Wissenschaft. Thanks to Sabrina Trapmann for data collection assistance.

3.1 Introduction

The sex-specific differential prediction of academic performance by scholastic aptitude tests has been studied extensively in US-American samples. Research indicates a small but persistent under-prediction of women's academic performance, at least for undergraduates (Young & Kobrin, 2001). The present study broadens the geographic and cultural scope by assessing the fairness of cognitive ability tests in three German samples.

Only two studies of differential prediction in Germany have been published so far, none of them peer-reviewed. The admission test to medical college (Test für Medizinische Studiengänge, TMS) shows no tangible differential prediction of two-year exam performance ($n = 19,561$); neither does the composite score of test and high school grade point average (HSGPA) (Nauels & Meyer, 1997). A custom-built law school admission test underpredicts women's subsequent bachelor grade point average in the first of two cohorts ($n = 63$ and $n = 91$, respectively); the composite score of test score, HSGPA, and an oral presentation shows no differential prediction (Dlugosch, 2005). Unfortunately, the predictor variables were not centered in these studies, so these effects may be artifacts (Schmidt & Hunter, 1982).

Following the recommendations of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), we use moderated multiple regression (MMR) analysis to assess differential prediction. The analysis of verbal and numeric subtests allows us to explore the relationship between underlying constructs and differential prediction, and to infer possible remedies against sex-specific discrimination in student selection.

3.2 Method

3.2.1 Sample 1

The first sample contains 116 freshman students of business administration at a German university, who took part in testing sessions three months after

beginning their studies. They received a monetary reimbursement and individual feedback in return. Cognitive ability was measured by the Berliner Intelligenzstruktur-Test (BIS) (Jäger, Süß, & Beauducel, 1997), a widely used German intelligence test that provides subscales for numeric and verbal abilities. Items assessing reasoning, memory, and speed were employed. Subtest reliabilities (Cronbach's α) range from .75 to .89. Participants also reported their HSGPA. Their subsequent college grade point average (CGPA) was obtained from the office of academic affairs one year later ($n = 87$).

3.2.2 Sample 2

A total of 914 students were initially recruited for Sample 2 at a German university. Their fields of study included agricultural science, biology, nutritional science, communication science, food technology, food chemistry, economic education, and economics. Over three quarters of the participants were either freshman or sophomore students. They received individual feedback after six weeks on request. Book vouchers were raffled among all participants. The longitudinal study design along with the exclusion of non-native speakers led to some attrition, leaving 728 persons in the final sample. Predictor variables were assessed in three months after the start of university. Grades were obtained from the office of academic affairs two years later. In this sample, CGPA pertains to the first two years of study, independent of a student's actual year in college. Again, the items from the BIS (Jäger et al., 1997) were used to assess numeric and verbal reasoning. All tasks of the respective scales were presented in a random order.

3.2.3 Sample 3

Deviating from the previous two samples, Sample 3 contains cross-sectional data. The cognitive ability test used in this sample was developed as a guidance tool for a German student counseling homepage (Hell, Päßler, & Schuler, 2009). It is also based on the Berliner Intelligenzstruktur model and contains (among others) subscales measuring verbal (e. g., word analogies, sentence completion, and antonyms) and numeric (e. g., number sequence,

rule-of-three problems, and arithmetic problems) abilities, respectively. Each subtest has reliability (Cronbach's α) of .73. CGPA pertains either to the main exam after the first year (Orientierungsprüfung), to the GPA of the first two years, or—when both were available—to the average of both. Students from colleges in Southern Germany were encouraged to participate by cooperating faculties and in some cases compensated with credits. Out of 3,170 persons who completed both the verbal and the numerical ability test sincerely, 1,801 native speakers provided their HSPGA, CGPA, and field of study and were subsequently included in Sample 3.

3.2.4 Data Analysis

The same statistical analysis is performed for each of the three samples. Numeric test scores, verbal test scores, and HSGPA are used as predictor variables. CGPA is the criterion variable. All GPAs are reversed in the analysis to mirror the American grading system, where higher grades indicate a better performance. Test scores are standardized to facilitate the comparison within and between samples and to avoid the misinterpretation of intercept differences in MMR (Schmidt & Hunter, 1982). CGPAs are standardized within fields of study to control for possible differences in grading leniency (Berry & Sackett, 2009).

Descriptive statistics are reported along with tests and effect sizes for sex differences on all variables, as recommended by Meade and Tonidandel (2010). Validity coefficients were calculated for each predictor as well as the combination of all predictors. Differential prediction was assessed by testing for equal slopes (group-test interaction) and equal intercepts (group dummy main effect) in MMR. In order to assess the extent of the predictive bias, we provide η^2 as effect size. A confidence level of $\alpha = .05$ is adopted for all tests.

3.3 Results

Descriptive statistics of the variables are listed in Table 3.1. The correlation coefficients in Table 3.2 indicate the criterion validity of each predictor by sample and sex. Sample 2 shows lower validities for the cognitive test scores

in comparison to HSGPA. HSGPA appears to be a less potent predictor in Sample 3, although it still trumps the ability subtests. The validities of test scores and HSGPA combined, on the other hand, are well within the range of previous studies (e. g., Berry & Sackett, 2009).

Differential prediction results are listed in Table 3.3 along with the corresponding regression equations. None of the MMR tests achieves significance in Sample 1. Still, the effect size for numeric test scores is the most pronounced in the present study, suggesting a small amount of under-prediction of women's academic performance. The MMR analysis of Sample 2 yields significant effects when HSGPA is included in the prediction. There is a very small amount of over-prediction of women's academic performance (i. e., unequal intercepts). The significant results in Sample 3 pertain to the numeric test score and to the test sum score. The effect sizes indicate a small amount of under-prediction of women's CGPAs.

3.4 Discussion

All sex-related tests of slopes and most tests of intercepts do not achieve significance. The ability tests under-predict women's CGPA significantly in Sample 1 and in Sample 3. The effect is small and in line with previous studies on differential prediction (Young & Kobrin, 2001), suggesting a slight bias against female test takers. The under-prediction of women's academic performance is reduced when the combination of test score and HSGPA is used as predictor. The over-estimation of men's academic performance is more pronounced when numeric ability scores rather than verbal ability scores are used in the prediction model. This discrepancy mirrors findings of US-American research (e. g., B. F. Patterson et al., 2009). Sample 2 shows no differential prediction for the ability tests. Here, using HSGPA as predictor leads to the under-prediction of men's academic performance. Consequently, the composite model is also prone to under-estimate men's CGPA. Again, the corresponding effect sizes are small.

3.4.1 Limitations

Each sample has its own limitations. Sample 1 is rather small; Sample 2 yields unequal error variances for predictions that contain HSGPA; and Sample 3 has a cross-sectional design and uses a custom-built test instrument. On the other hand, this heterogeneity can also be seen as a strength of our study as we use different tests and focus on different populations with different designs. The combined results provide a detailed account of differential prediction in German colleges.

HSGPA was an admission criterion in Sample 1 and Sample 2, and in some cases of Sample 3. We did not correct for possible restrictions of range, partly because they apply only to a subset of students, and partly because data from unrestricted samples are scarce. This might lead to a slight underestimation of the validity coefficients and the amount of predictive bias, but it is unlikely to skew the sex-specific differential prediction.

Finally, the testing situation in the present study differs from high-stakes admission tests. Additional differential effects might occur in more competitive setting. Still, our results are in line with the existing research on differential prediction, which suggests that US-American findings also apply to student selection procedures in Germany.

3.4.2 Conclusion

Differential prediction does not fundamentally undermine the fairness in ability test use in this German scenario. Moreover, our results suggest that one can possibly design a predictor that has minimal bias by combining HSGPA and the appropriate intelligence domains. This notion is supported by SAT studies that found a larger predictive bias against women for mathematical SAT parts than for verbal SAT parts (e. g., B. F. Patterson et al., 2009). Given the small sex-specific effect sizes, actions to reduce differential prediction should be subtle and aimed at obtaining a fair balance.

Table 3.1: Descriptive statistics of each sample by gender (means and standard deviations) and pair-wise comparison. Test scores are standardized. GPAs use reversed German scoring (i. e., higher numbers indicate better grades).

Sample	Variable	Men	n	Women	n	d
Sample 1	Numeric	0.32 (1.10)	46	−0.36 (0.74)	41	0.68**
	Verbal	−0.04 (0.99)	46	0.04 (1.02)	41	−0.08
	Numeric + Verbal	0.18 (1.10)	46	−0.20 (0.84)	41	0.37°
	HSGPA	2.67 (0.65)	46	2.79 (0.65)	41	−0.19
	CGPA	1.95 (0.80)	46	2.10 (0.79)	41	−0.19
Sample 2	Numeric	0.29 (1.01)	293	−0.20 (0.94)	435	0.49***
	Verbal	0.11 (1.02)	293	−0.08 (0.98)	435	0.19*
	Numeric + Verbal	0.24 (1.00)	293	−0.16 (0.97)	435	0.40***
	HSGPA	2.73 (0.58)	293	2.92 (0.58)	435	−0.32***
	CGPA	2.14 (0.78)	293	2.33 (0.87)	435	0.09
Sample 3	Numeric	0.42 (0.79)	768	−0.31 (1.02)	1,033	0.73***
	Verbal	0.12 (0.94)	768	−0.09 (1.04)	1,033	0.21***
	Numeric + Verbal	0.32 (0.87)	768	−0.24 (1,02)	1,033	0.55***
	HSGPA	2.94 (0.61)	768	2.97 (0.60)	1,033	−0.05
	CGPA	2.75 (0.70)	768	2.80 (0.68)	1,033	−0.05

Notes: ° $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Table 3.2: Validity of each predictor and selected combinations; CGPA is used as criterion. Test is the sum score of numeric and verbal. Composite is test score together with HSGPA; r from a linear regression with these two predictors is reported here.

Sample	Variable	CGPA	Numeric	Verbal	Test	HSGPA	Composite
Sample 1	CGPA	–	.37	.29	.36	.44	.45
	Numeric	.19	–	.77	.96	.59	.96
	Verbal	.26	.50	–	.92	.60	.92
	Test	.26	.84	.89	–	.63	[a]
	HSGPA	.55	.22	.39	.36	–	[a]
	Composite	.56	.85	.89	[a]	[a]	–
Sample 1	CGPA	–	.14	.18	.19	.48	.48
	Numeric	.11	–	.37	.83	.21	.83
	Verbal	.17	.42	–	.83	.37	.83
	Test	.16	.84	.85	–	.35	[a]
	HSGPA	.43	.16	.38	.33	–	[a]
	Composite	.43	.85	.86	[a]	[a]	–
Sample 2	CGPA	–	.16	.20	.22	.41	.41
	Numeric	.19	–	.39	.79	.34	.79
	Verbal	.24	.37	–	.88	.37	.88
	Test	.26	.81	.85	–	.43	[a]
	HSGPA	.39	.42	.42	.51	–	[a]
	Composite	.39	.81	.85	[a]	[a]	–

Notes: Correlations for men are presented above the diagonal, and correlations for women are presented below the diagonal. See Table 3.1 for sample sizes.

[a] Not computed because predictors and outcome overlap.

Table 3.3: Differential prediction analysis with MMR equations (standard errors in brackets); outcome variable is standardized CGPA.

		Test of Slopes			Test of Intercepts			
Sample	Predictors	F	df	η^2	F	df	η^2	Equation
1	Numeric	0.17	1,83	.002	3.11	1,83	.034°	.19(.164) + .24(.20)N − .39(.22)M + .10(.24)N*M
	Verbal	0.06	1,83	.000	0.62	1,83	.008	.09(.148) + .25(.15)V − .17(.20)M + .05(.21)V*M
	Test	0.02	1,83	.000	2.22	1,83	.024	.16(.151) + .30(.18)T − .31(.21)M + .03(.22)T*M
	HSGPA	0.16	1,83	.001	0.28	1,83	.003	.05(.135) + .52(.14)H − .10(.19)M − .08(.19)H*M
	Composite	0.20	2,81	.004	0.58	1,81	.005	.07(.142) + .09(.18)T + .49(.15)H − .15(.20)M + .04(.23)T*M − .13(.22)H*M
2	Numeric	0.04	1,724	.000	0.12	1,724	.000	−.01(.05) + .11(.05)N + .03(.08)M + .02(.08)N*M
	Verbal	0.01	1,724	.000	0.58	1,724	.001	−.02(.05) + .17(.05)V + .06(.07)M − .01(.07)V*M
	Test	0.02	1,724	.000	0.05	1,724	.000	−.01(.05) + .17(.05)T + .02(.08)M + .01(.08)T*M
	HSGPA	0.04	1,724	.000	11.84	1,724	.013**	−.09(.04) + .45(.04)H + .23(.07)M + .01(.07)H*M
	Composite	0.02	2,722	.000	9.72	1,722	.011**	−.09(.04) + .03(.05)T + .44(.05)H + .22(.07)M + .00(.07)T*M + .01(.07)H*M
3	Numeric	0.27	1,1797	.000	14.21	1,1797	.008***	.08(.03) + .18(.05)N − .19(.05)M + .03(.05)N*M
	Verbal	0.02	1,1797	.000	4.31	1,1797	.002*	.04(.03) + .23(.03)V − .10(.05)M − .01(.05)V*M
	Test	0.03	1,1797	.000	15.39	1,1797	.008***	.08(.03) + .25(.03)T − .19(.05)M + .01(.05)T*M
	HSGPA	0.42	1,1797	.000	0.52	1,1797	.000	.01(.03) + .38(.03)H − .03(.04)M + .03(.04)H*M
	Composite	0.44	2,1795	.000	2.37	1,1795	.001	.03(.03) + .08(.03)T + .34(.03)H − .07(.05)M − .01(.05)T*M + .05(.05)H*M

Notes: N = Numeric, V = Verbal, T = Test, H = HSGPA, M = Male gender dummy; ° $p < .10$, $*p < .05$, ** $p < .01$, *** $p < .001$

4 Study 2: Women and Men Tend to Use Different Narrow Abilities in Tests of Scholastic Aptitude

Pre-print manuscript. See Section 1.2 for author and publication details[15].

Abstract

This study explores how intelligence facets are mapped in tests of scholastic aptitude for women and men. Intelligence test scores and academic aptitude test scores from freshman students in science ($n = 284$) and economics ($n = 358$) as well as subsequent grades from their first year in college were used to analyze structural equation models. The direct influence of intelligence facets on academic performance is completely mediated by academic aptitude test scores. Numeric abilities dominate the aptitude test's predictive power. Verbal abilities are important in science but not in economics. The validities of some narrow abilities differ between women and men. The findings further suggest that intelligence facets are covered sufficiently by admission tests.

Keywords: college admission, intelligence, sex differences, academic achievement, criterion validity

4.1 Introduction

Standardized achievement tests are a crucial tool in most selection situations. Their main task is to predict the future performance of applicants. In higher education, tests of scholastic aptitude are preferred over measures of general intelligence. Tests like the SAT and the ACT are supposed to assess not just

[15]This work was supported by a grant by the German Federal Ministry of Education and Research [grant agreement number: 01FP0930] and the European Social Fund of the European Union awarded to Benedikt Hell. Parts of the raw data used in the present study have been previously analyzed and published in a different context in F. Fischer, Schult, and Hell (2012). Sex differences in secondary school success: Why female students perform better. *European Journal of Psychology of Education*, doi: 10.1007/s10212-012-0127-4.

general cognitive ability, but particularly scholastic aptitude. The concept of scholastic aptitude is diffuse at best (F. Patterson & Ferguson, 2010). The present study shows how facets of cognitive ability are mapped on a set of German scholastic aptitude tests and how these tests mediate the predictive validity of intelligence.

The predictive power of college admission tests relies heavily on the same underlying psychological construct: intelligence. The correlation between IQ and admission test scores ranges between .5 and .8 (Frey & Detterman, 2004; K. A. Koenig et al., 2008). Some researchers even go as far as using the SAT as a measure of general intelligence (Jackson & Rushton, 2006) and calling it a "[test] of general ability" (Carroll, 1993, p. 704), although it was not designed for this purpose. The ACT homepage insists that "[t]he ACT is not an aptitude or an IQ test" (ACT, 2013), but a curriculum-based assessment. Still, the correlation between ACT and intelligence lies between .6 and .7, and the correlation between ACT and SAT scores is large ($> .8$) (K. A. Koenig et al., 2008). Also, SAT scores correlate with up to .7 with intelligence measures like the Wonderlic Personnel Test (Coyle & Pillow, 2008). All these correlations are usually even larger when corrected for restriction of range (Sackett & Yang, 2000).

General college admission tests usually have a large predictive validity when it comes to predicting academic achievement across large samples (Berry & Sackett, 2009). They are supposed to measure specific aptitudes that are crucial for educational success, but they usually probe general mental ability to a certain extent (F. Patterson & Ferguson, 2010; Coyle & Pillow, 2008). The outcome of interest in most validity studies is first-year college grade point average (CGPA); the relationship between test performance and academic success remains remarkably stable for longer time lags between test and criterion assessment (Sackett et al., 2008). Other success criteria include satisfaction ratings, retention, and subsequent chances on the job market (Stemler, 2012). Validity coefficients of SAT subtests critical reading (SAT-CR), mathematics (SAT-M) and writing (SAT-W) are quite similar: Correlations with first-year CGPA range between .34 and .36 for women and between .28 and .31 for men (B. F. Patterson et al., 2009). For the Graduate

Management Admission Test (GMAT), the difference between the validity coefficients of GMAT-Verbal and GMAT-Quantitative is small and its sign depends on the criterion (first-year grade point average, graduate grade point average, or persistence; Kuncel et al., 2007).

Subject-specific admission tests like the GMAT aim at tailored predictions for particular academic fields. Their predictive validity is moderate (e. g., in economics) to large (e. g., in medicine) according to meta-analytic studies (Donnon et al., 2007; Hell et al., 2007; Oh et al., 2008).

The predictive validity of college admission tests cannot be fully explained by general intelligence (Coyle & Pillow, 2008), but – maybe more interestingly – it is also unclear which aspects of intelligence are reflected in measures of scholastic aptitude. And does intelligence predict academic achievement after removing scholastic aptitude test performance?

4.1.1 Facets of Intelligence in Admission Testing

The presence of an overarching factor of general intelligence, *g*, is widely accepted (Jensen, 1998; McGrew, 2009). Still, our understanding of intelligence can be enhanced by studying rotated factor solutions and also multilevel-factor models (Carroll, 1993; Nisbett et al., 2012).

The dominant intelligence factor in most college admission tests is reasoning (Atkinson & Geiser, 2009), which is also known as fluid intelligence (Gf; Carroll, 1993) and fluid reasoning (see McGrew, 2009). It facilitates the creation of secondary abilities that are crucial for educational success (Geary, 2010). Consequently, reasoning items feature prominently in established college admission tests (Zwick, 2007).

Despite the close relationship between intelligence and admission testing, there have been few studies that went beyond simple correlations between *g* factor and test score. Bivariate correlations depend partly on the intelligence test used: Tests of general intelligence like the Wonderlic Personnel Test show very high correlations with admission test scores, as do subtests of arithmetic reasoning and verbal synonyms and antonyms ($r > .6$) (Coyle & Pillow, 2008; K. A. Koenig et al., 2008). Correlations are lower for figural

tests and for tests of coding speed ($r < .4$) (Coyle & Pillow, 2008).

4.1.2 Sex Differences in Admission Testing, IQ and CGPA

A comprehensive overview of sex differences for adolescents and adults in Europe and North America is provided by Ellis and colleagues (2008): Men tend to get higher scores than women in college admission exams. On the other hand, there is no pronounced sex difference in general tests of language or verbal ability. When differences are found, women tend to get higher scores on language-related reasoning subtests. At the level of the overarching g factor, sex differences are usually very small, although some findings suggest that men's IQs are slightly higher in early adulthood (e. g., Irwing & Lynn, 2005; Lynn & Kanazawa, 2011). All these constructs may serve as promising predictors of academic performance, which is usually operationalized via grades. First-year CGPA is used as criterion most frequently; here, male students are on average slightly better than female students (Mattern & Patterson, 2013). The present study explores how different ability facets work together in explaining academic performance and whether there are different validity patterns for women and men.

In most studies of college admission tests, validity coefficients are larger for women than for men (Young & Kobrin, 2001). A common explanation for this difference is a larger variance in men's study behaviors that diminishes the predictive power of their test scores (Zwick, 2007). Even when differences in cognitive ability are small, differential validity may also exist with regard to the relationship between intelligence facets and admission test scores. Men might apply their mental resources in other ways than women due to differences in self-confidence and other personality traits (Else-Quest, Hyde, & Linn, 2010). As in the ability domain of spatial reasoning, women and men possibly use different ability patterns to solve test items (cf. Kaufman, 2007). Identifying facet-specific differential validity should help to understand sex differences in admission testing better.

4.1.3 Aim of the Present Study

We look at the loadings of intelligence facets on scholastic aptitude test performance (ATP) and assess the predictive validity of these constructs with regard to subsequent CGPA. Using structural equation models, we explore how the predictive power of intelligence is mediated by scholastic aptitude and to what degree specific intelligence facets are insufficiently covered by the scholastic aptitude tests. A secondary aim of the present study is to test for sex differences in order to investigate the role of intelligence facets in the differential validity of the scholastic aptitude tests.

4.2 Method

4.2.1 Sample and Study Design

During their first weeks at university, 670 science and economics freshman students (52.7 % women; age in years: $\bar{x} = 20.3$, $SD = 1.9$) from two German universities were administered tests of scholastic aptitude and intelligence. These particular fields of study were chosen because here admission is often competitive. Also, costs for science studies are above average, making drop-outs a larger financial burden than in most other fields. Participation was voluntary. The samples constitute about 50 % of the respective freshman populations. All instructors were female to reduce the influence of stereotype threat for women (Walton & Spencer, 2009).

A complete list of college grade points was obtained from the office of student administration one year after the initial testing session. CGPAs were then calculated for each semester. We modeled latent ability factors based on manifest semester CGPAs in order to increase the reliability of the scholastic aptitude measurement (Deary et al., 2007). German CGPAs have been reversed so that here 4 indicates the best grade and 0 indicates the lowest grade.

4.2.2 Instruments

The intelligence test used in the present study, the I-S-T 2000 R (Liepmann, Beauducel, Brocke, & Amthauer, 2007), is based on the *Berlin model of intelligence structure*, which distinguishes between two modalities: operations (speed, memory, creativity, reasoning) and contents (verbal, numeric, figural; Jäger, 1984). We used two reasoning subtests per content factor, choosing the ones that were most reliable according to the manual: verbal analogies, verbal similarities, calculations, number series, abstract pieces, and cubes. About half of the participants worked Form A whereas the rest worked Form B, which contains the same items as Form A but in a different order and – when possible – with shuffled distractors.

The tests of subject-specific scholastic aptitude (Studienfeldbezogene Beratungs-Testserie, SFBT) were provided by the federal employment office. Although they are usually used for counseling purposes, they closely resemble actual college admission tests. According to the manuals, no prior knowledge beyond high school knowledge is required to take the tests. The science subsample was administered the test specifically designed for prospective science students and the economics subsample was administered the test specifically designed for prospective economics students. Each test had two parts. The science test started with 20 items regarding the basic understanding of science problems, followed by 20 items covering scientific diagrams and tables (50 minutes each). The economics test started with 20 items that deal with calculations of economic models (60 minutes), followed by 20 items covering economic diagrams and tables (55 minutes). The response format was multiple choice with 4 (science) and 5 (economics) options, respectively. The manuals (Bundesagentur für Arbeit, 2004a, 2004b) report moderate (uncorrected) validity coefficients for CGPA after two years ($.21 < r < .32$ for science, $.27 < r < .37$ for economics). Possible scores range from 0 to 40. For the present subsamples, reliabilities are $\alpha = .83$ (science) and $\alpha = .90$ (economics), respectively.

4.2.3 Data Analysis

Cases whose self-reported student ID did not match the administration data were excluded, leaving 642 students in the final analysis sample: 284 science students (51 % women) and 358 economics students (53 % women).

Structural equation modeling (SEM) was performed with the raw data using maximum likelihood estimation with missing values (mlmv) in Stata 12.1. Separate models were estimated for science and economics students, respectively. In order to identify sex differences, we estimated the structural paths simultaneously for men and women within each subsample using multiple-group SEM. Means of latent variables were also allowed to vary by sex because previous research suggests group differences on test scores (B. F. Patterson et al., 2009; Steinmayr, Beauducel, & Spinath, 2010). Loadings on manifest indicators, error variances, and covariances between exogenous variables were constrained to be equal across sex. Wald tests were performed to probe group differences. All tests for differences are two-tailed. Model fit was assessed using the model χ^2 (Jöreskog, 1969), the Comparative Fit Index (CFI), and the Root Mean Square Error of Approximation (RMSEA). CFI values above .95 indicated an adequate model fit (Hu & Bentler, 1999), whereas RMSEA values below .08 were considered acceptable (Browne & Cudeck, 1993).

Comparison of the Bayesian information criterion (BIC) and Likehood Ratio (LR) tests were used to test whether the measurement of latent variables can be constrained to be τ-equivalent (equal loadings – a mandatory constraint for two-indicator measurement models if there was no multiple-group design) or equivalent (equal loadings and equal error variances).

The theoretical model is shown in Figure 4.1. It contains the loadings of the reasoning facets on ATP as well as the full mediation model with regard to CGPA. The predictive validities of the SFBTs are modeled with direct effects on CGPA. Direct effects from reasoning facets on CGPA are also allowed to test whether the predictive power of intelligence measures is mediated partially or completely.

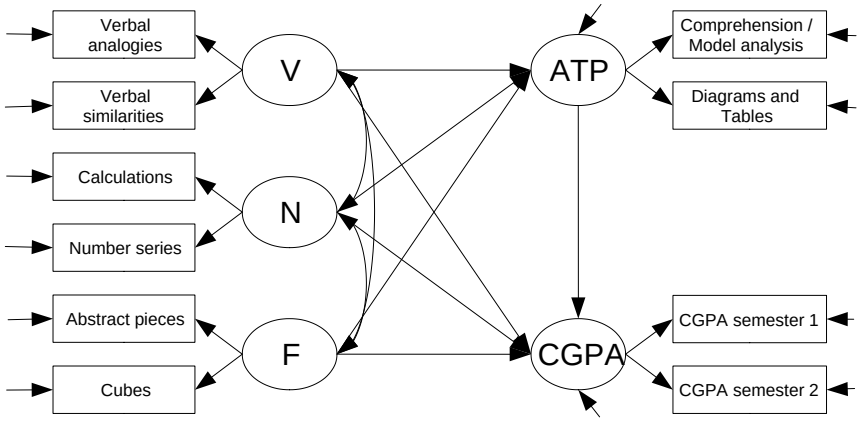


Figure 4.1: Theoretical SEM with intelligence facets (V = verbal, N = numeric, F = figural), scholastic aptitude test performance (ATP), and first-year college grade point average (CGPA)

4.3 Results

4.3.1 Descriptive Statistics

Sample statistics of the variables of interest are listed by sex and field of study in Table 4.1. Table 4.2 shows the correlations between all variables for men and women. Both tables contain data from cases without any missing values. First-semester CGPA had the most missing values ($n = 66$); second-semester CGPA had 55 missing values. Men outperformed women on most subtest scores, but did not earn better college grades.

4.3.2 Structural Equation Models

Given the high correlations between the three intelligence variables, the main models were compared to a simpler SEM with only one latent intelligence factor (g). Fit indices suggest that the more complex models with three facets were more adequate than the g factor model (see Table 4.3). The models with τ-equivalent measurement yielded the best (i. e., lowest) BIC values in each subsample. LR tests point towards these models, as well. Table 4.3 shows the model comparisons.

The main model (with τ-equivalent indicators) for science students is

shown in Figure 4.2. The model fit was sufficient ($\chi^2 = 119.26$, $p = .006$; CFI = .954; RMSEA = .055). It could be improved by relaxing several constraints according to the modification indices, but for the sake of simplicity and in order to avoid an overfit only the main model is presented here despite the somewhat unsatisfactory χ^2 statistics.

The path coefficients did not differ significantly between men and women. Some absolute differences that were rather large suggest that the lack of significant difference could be at least partially a statistical power problem caused by the sample size of less than 300 (Saris & Satorra, 1993).

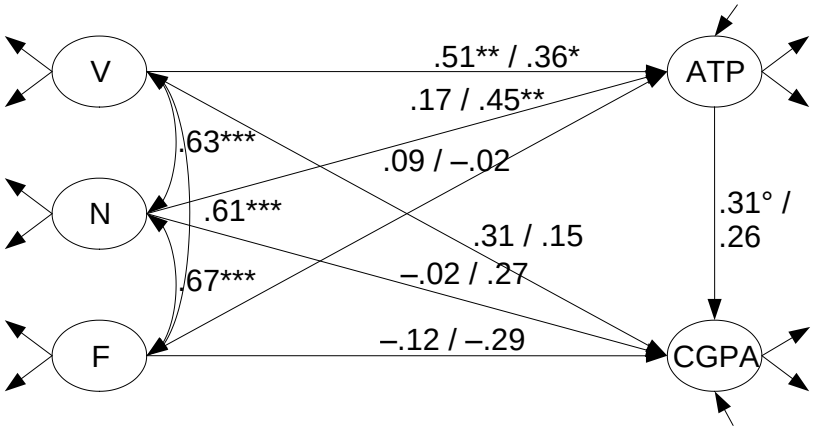


Figure 4.2: Standardized SEM path coefficients for the science model for women / men ($^\circ p < .10$, $^*p < .05$, $^{**}p < .01$, $^{***}p < .001$, $n = 284$)

The main model (with τ-equivalent indicators) for economics students is shown in Figure 4.3. Again, the model fit was predominantly satisfactory ($\chi^2 = 143.02$, $p < .001$; CFI = .953; RMSEA = .064). The path coefficient from numeric reasoning to ATP was significantly larger for men ($p = .029$), whereas the path coefficient from figural reasoning to ATP was marginally larger for women ($p = .053$).

There were two indirect effects that approach significance – V via ATP on CGPA (women, science; $p = .099$) and N via ATP on CGPA (men, economics; $p = .077$) – and only one that is significant – F via ATP on CGPA (women, economics; $p = .042$).

Table 4.1: Descriptive statistics for women and men and group comparisons (positive d values indicate higher scores for women)

		Women		Men			
Sample	Variable	M	SD	M	SD	p	Cohen's d
Science	Verbal analogies	12.7	2.6	13.6	2.7	.005	−.34
	Verbal similarities	12.1	3.0	12.8	2.7	.041	−.24
	Calculations	13.2	3.6	15.3	3.4	.000	−.60
	Number series	13.7	4.6	15.7	4.3	.000	−.46
	Abstract pieces	12.2	4.0	13.7	3.7	.001	−.39
	Cubes	11.6	4.2	13.2	4.1	.001	−.41
	Comprehension	10.1	3.0	12.4	2.9	.000	−.78
	Diagrams and Tables	12.1	3.1	14.7	3.0	.000	−.86
	CGPA semester 1	2.3	0.87	2.4	0.83	.22	−.17
	CGPA semester 2	2.4	0.86	2.3	0.95	.47	.09
Economics	Verbal analogies	10.8	2.9	11.6	3.2	.016	−.26
	Verbal similarities	10.8	3.2	10.8	3.6	.97	−.00
	Calculations	11.8	3.7	13.3	3.9	.000	−.40
	Number series	12.6	4.8	15.0	4.5	.000	−.50
	Abstract pieces	10.1	3.6	10.6	4.0	.17	−.15
	Cubes	10.3	4.0	11.1	4.5	.074	−.19
	Model analysis	8.0	3.4	9.9	4.2	.000	−.50
	Diagrams and Tables	8.3	3.0	10.2	3.2	.000	−.64
	CGPA semester 1	2.1	0.87	2.2	0.89	.29	−.11
	CGPA semester 2	1.8	0.96	2.0	0.93	.29	−.12

Note. Science subsample: $123 < n(\text{women}) < 146$; $95 < n(\text{men}) < 138$. Economics subsample: $178 < n(\text{women}) < 189$; $156 < n(\text{men}) < 169$.

Table 4.2: Correlation matrices for the variables of interest

Science ($p < .05$ for all $r > .15$ except †)										
	1	2	3	4	5	6	7	8	9	10
1. Verbal analogies	–	.45	.32	.32	.43	.27	.40	.34	.22	.19
2. Verbal similarities	.48	–	.37	.21	.24	.15	.29	.25	.14	.08
3. Calculations	.44	.29	–	.51	.38	.35	.37	.42	.15	.25
4. Number series	.29	.26	.57	–	.26	.35	.33	.38	.20†	.19
5. Abstract pieces	.31	.41	.45	.38	–	.56	.28	.41	.16	.13
6. Cubes	.14	.25	.29	.29	.39	–	.12	.19	.04	−.03
7. Comprehension	.42	.33	.29	.29	.22	.14	–	.55	.33	.19
8. Diagrams and Tables	.31	.27	.38	.28	.43	.20	.55	–	.26	.24
9. CGPA semester 1	.31	.26	.24	.10	.27	.07	.18	.29	–	.80
10. CGPA semester 2	.20	.19	.21	.02	.14	−.12	.30	.35	.65	–
Economics ($p < .05$ for all $r > .14$)										
	1	2	3	4	5	6	7	8	9	10
1. Verbal analogies	–	.54	.44	.43	.38	.28	.25	.34	.21	.10
2. Verbal similarities	.61	–	.40	.35	.42	.26	.16	.32	.18	.05
3. Calculations	.50	.45	–	.64	.50	.29	.42	.43	.37	.39
4. Number series	.46	.43	.58	–	.37	.29	.43	.42	.21	.26
5. Abstract pieces	.39	.43	.36	.47	–	.50	.29	.31	.18	.10
6. Cubes	.36	.32	.38	.52	.44	–	.18	.20	.12	.07
7. Model analysis	.24	.14	.40	.28	.34	.21	–	.50	.31	.29
8. Diagrams and Tables	.37	.28	.33	.37	.41	.31	.37	–	.34	.37
9. CGPA semester 1	.31	.31	.38	.25	.37	.09	.40	.41	–	.79
10. CGPA semester 2	.13	.08	.21	.18	.18	−.00	.32	.32	.77	–

Note. Correlations for men are presented above the diagonals;
correlations for women are presented below the diagonals.
Science subsample: $113 < n(\text{women}) < 146$; $88 < n(\text{men}) < 138$;
Economics subsample: $176 < n(\text{women}) < 189$, $156 < n(\text{men}) < 169$.

Table 4.3: SEM comparisons for facets and g factor

Model	χ^2	df	$\Delta\chi^2$	df	p	CFI	RMSEA	BIC
Science g factor	177.57	91				.890	.082	12662
Science facets	112.35	78				.956	.056	12670
Science facets (τ-equivalent)	119.26	83	6.92	5	.23	.954	.055	12649
Science facets (equivalent)	165.19	88	45.92	5	$< .001$	.902	.079	12667
Economics g factor	219.28	91				.900	.089	16446
Economics facets	133.27	78				.957	.063	16437
Economics facets (τ-equivalent)	143.02	83	9.75	5	.08	.953	.064	16417
Economics facets (equivalent)	231.42	88	65.22	5	$< .001$	.889	.095	16476

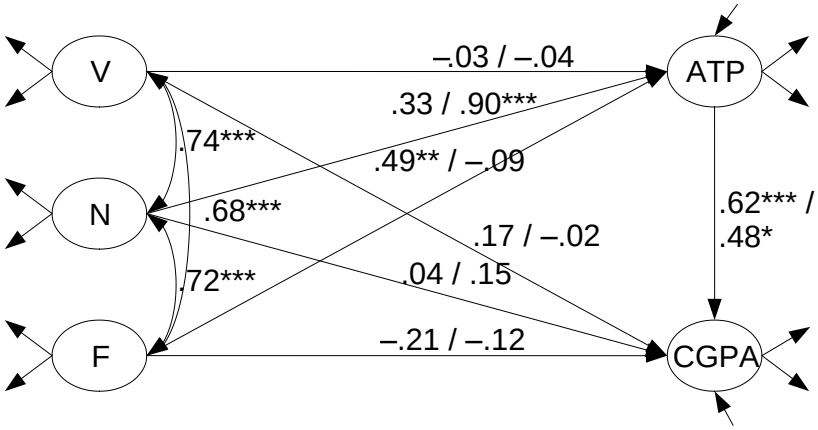


Figure 4.3: Standardized SEM path coefficients for the economics model for women / men ($^\circ p < .10$, $^*p < .05$, $^{**}p < .01$, $^{***}p < .001$, $n = 358$)

4.4 Discussion

Our data suggests that it is useful to look beyond g when it comes to subject-specific tests of scholastic aptitude. The relationship between intelligence facets and ATP depends on the field of study and, at least in economics, also on sex.

4.4.1 Intelligence Facets Matter

Both main models suggest that intelligence facets matter when it comes to ATP, but the relationships vary between fields of study and sex.

Verbal reasoning loads on ATP in the science subsample. It seems like verbal skills play an important role in comprehending the scholastic aptitude test items, which are more complex than the intelligence test items and contain verbose context information. Numeric reasoning is also relevant for science students, at least for male ones. This is not surprising, given the arithmetic nature of most items. Figural reasoning is the least relevant intelligence facet, although women appear to use numeric and figural skills jointly to some degree.

In the economics subsample, the role of verbal reasoning – which correlates highly with the other two intelligence facets and whose indicators have larger variances compared to the science subsample (each $p < .025$) – is negligible. Here it is numeric reasoning that shows the largest loadings on ATP. The path coefficient is significantly smaller for women, who appear to tap more into their figural skills instead.

The mean sex differences on the variables of interest resemble previous research findings (B. F. Patterson et al., 2009; Steinmayr et al., 2010) and data reported in the respective test manuals. Men tend to have higher scores on numeric and figural subtests as well as on the scholastic aptitude tests. There are no significant CGPA differences in our sample.

The slightly larger path coefficients for women from ATP to CGPA are in line with previous findings (Zwick, 2007). Unfortunately, the sample sizes appear to be too small to identify this small sex difference.

The differential relationship of intelligence facets with ATP suggests that men resort to an arithmetically inclined set of cognitive skills for their academic performance. Their scholastic aptitude test scores reflect aspects of intelligence that are rather homogeneous with a strong focus on numerical reasoning. Women, on the other hand, appear to have a somewhat different arsenal of deployable skills at their disposal for solving scholastic aptitude tests. Consequently, their test scores may cover constructs that are relevant for academic achievement in a more diverse and possibly more precise way, leading to differential validities with higher coefficients for women. These findings are in line with a social/personality explanation of sex differences in scholastic aptitude test score (Hannon, 2012).

The relationship between intelligence facets and college performance is spurious when ATP included as mediator. All direct effects on CGPA fail to achieve significance. The notion that women use verbal skills that are not reflected in their ATP (science: $\gamma_{(\mathrm{V,ATP})} = .31$, economics: $\gamma_{(\mathrm{V,ATP})} = .17$) is merely speculative and needs to be replicated in larger samples.

The aptitude test scores show more consistent validity coefficients than the intelligence subtests (see Table 4.2). Cognitive ability cannot fully explain ATP (see also Coyle & Pillow, 2008). Further inspection of correlations and

modification indices also suggests that figural abilities that are assessed in the cubes subtests cannot explain ATP and CGPA; instead, they seem to impair women's academic achievement. The cubes task requires three-dimensional rotation, in which men tend to outperform women (Nisbett et al., 2012). Another possible explanation for these contra-intuitive findings is fatigue. Participants had been solving items for over three hours (including a short break after the scholastic aptitude test) when the last reasoning subtest (i. e., cubes) came on. The lack of a high-stakes testing situation may have led to a decline in smarter women's motivation.

Previous research suggested that figural intelligence measures correlate only moderately with admission test scores (Coyle & Pillow, 2008). The figural facet has little to no explanatory power in terms of ATP in the present models. For men, there appears to be no effect from F on CGPA beyond the indirect effect via ATP, so constraining the direct path from F to CGPA to zero improves the model. But given the risk of overfitting the models to the data, and also the aptitude test at hand, modifications of the main models are likely to reduce their reproducibility in other samples and other fields of study.

4.4.2 Limitations

Statistical power There were some participants in the science subsample (36 %) and many in the economics subsample (84 %) who did not answer all aptitude test items in the allotted time. The I-S-T 2000 R is also a speeded test. Test scores therefore tend to reflect a speed component in addition to reasoning ability, which might have a smaller predictive validity than reasoning (cf. Coyle & Pillow, 2008). This dilutes the construct validity of our measures. Still, regular college admission tests do have a time limit, a fact that is reflected in our study design.

Using just two indicators for each latent variable, the present model leaves little room to model detailed sex differences without risking identification issues. Future studies should strive to obtain more clear-cut, possibly narrower intelligence measurements.

Discerning test effects and sampling effects The comparison of the science model and the economics model is difficult, because differential effects could be attributed to sample composition or to the fact that a different aptitude test was used in each sample. Both tests share surface properties, response format, item complexity and level of abstraction. Item content, however, is clearly aimed at the respective field of study.

Insufficiencies of scholastic aptitude tests The present study can provide only an incomplete answer to the question which relevant constructs are not or only partially covered by scholastic aptitude tests. The mediation analysis suggests that reasoning skills are sufficiently reflected in the tests. Given the complexity of course contents in college, non-cognitive skills are probably helpful in optimizing the way students employ their cognitive abilities. The most promising candidates are personality traits like conscientiousness and intellectual curiosity (von Stumm, Hell, & Chamorro-Premuzic, 2011).

Looking at item characteristics is another promising approach to studying the predictive validity of academic aptitude tests (Kobrin, Kim, & Sackett, 2012). The tests used in the present study contain rather homogeneous items that are relatively complex and require a moderate level of ability to abstract. This is in line with the tests' predictive validity. Still, item format does not vary enough to expand our analysis into this direction.

4.4.3 Conclusion

The most consistent findings are that scholastic aptitude tests are valid predictors of academic achievement (in the science subsample at least in terms of bivariate correlations) and that numeric reasoning shows particularly high loadings on ATP. The pronounced effect of verbal intelligence in the science subsample suggests that college admission in fields like biology, chemistry, and physics depends not only on mathematical skills but is also related to the ability to make sense of verbal information.

In economics, men appear to rely mainly on numeric skills whereas women appear to draw on a broader set of facets for their academic performance. The

rather homogeneous set of cognitive skills used by male students is a possible explanation for their diminished validity coefficient between the scholastic aptitude test and academic success with its inherent complexity. Validity coefficients appear to reflect different aspects of intelligence for men and for women to some degree. This result is in line with a cognitive explanation for differential validity.

5 Study 3: Prädiktoren des Berufserfolgs von Hochschulabsolventen: Befunde aus dem Sozio-Ökonomischen Panel

Post-print manuscript of Schult (2012), appearing courtesy of Pabst Science Publishers (`http://www.psychologie-aktuell.com/index.php?id=183`). See Section 1.1 for author and publication details[16].

Zusammenfassung

Mit Hilfe von repräsentativen Längsschnittprofilen ($154 \leq n \leq 589$) wird untersucht, wie weit Schulleistungen und Persönlichkeitseigenschaften mit dem Berufserfolg von Hochschulabsolventen zusammenhängen. Als einziger Big-Five-Faktor zeigt Gewissenhaftigkeit eine substanzielle prädiktive Validität hinsichtlich der Arbeitszufriedenheit zwei Jahre nach Ende des Studiums. Bei der Prognose des Einkommens sind fachliche Interessen, beruflicher Status und Mathematik-Schulnoten die besten Prädiktoren. Dabei kann die Lohnlücke zwischen Frauen und Männern durch sozio-ökonomische und motivationale Faktoren nicht vollständig erklärt werden. Einkommensunterschiede bei den Interessendimensionen nach Holland (1997) reflektieren geschlechtsspezifische Präferenzen bei der Studienfachwahl. Weiterhin zeigt sich, dass sich eine niedrige Ausprägung der Persönlichkeitseigenschaft Verträglichkeit bei Frauen, nicht jedoch bei Männern negativ auf das Gehalt auswirkt.

Schlüsselworte: Berufserfolg, Arbeitszufriedenheit, Geschlechtsunterschiede, prädiktive Validität, Big Five, Schulnoten

[16]Der Autor dankt Benedikt Hell, Franziska Fischer, Katja Päßler, Michael Dantlgraber und Patrick Theiner für hilfreiche Kommentare zu früheren Fassungen dieses Manuskripts.

Anmerkung: Aus Gründen der Lesbarkeit wird die männliche Form stellvertretend für Männer und Frauen verwendet; Geschlechtsunterschiede werden jeweils explizit als solche gekennzeichnet.

Abstract: Predicing Professional Success of College Graduates: Findings from the German Socio-Economic Panel

This study investigates how school achievement and personality traits relate to professional success, using representative data from the German Socio-Economic Panel ($154 \leq n \leq 589$). Conscientiousness is the only Big Five factor that has a substantial predictive validity with regard to job satisfaction two years after graduation from college. Personal interests, occupational status, and school grades in mathematics are the best predictors for income. The gender pay gap cannot be fully explained by socio-economic and motivational factors. The gap reflects gender-specific preferences for particular subjects (classified according to the dimensions of interest by Holland, 1997). Low scores on the personality trait agreeableness are associated with a decreased income for women, but not for men.

Keywords: vocational achievement, job satisfaction, sex differences, predictive validity, Big Five, school grades

5.1 Einleitung

Welche psychologischen Merkmale eignen sich zur Vorhersage des Berufserfolgs von Hochschulabsolventen? Diese Frage ist in der Personalauswahl ebenso relevant wie bei der Beratung von Jungakademikern. Es gibt bereits umfassende Meta-Analysen zur Prognose von Berufserfolg durch Persönlichkeitseigenschaften (Judge et al., 1999, 2002) und Bildungsleistungen (Roth et al., 1996; Roth & Clarke, 1998). Im Anschluss an einen Überblick über die bestehenden Arbeiten wird in der vorliegenden Studie anhand von repräsentativen Paneldaten untersucht, wie weit sich diese Befunde auf die Prognose des Berufserfolgs übertragen lassen.

5.1.1 Noten als Leistungsmaß

Aus der Perspektive der Personalauswahl stellt ein Hochschulabschluss in erster Linie eine Qualifikationsleistung dar. Der Abschluss zertifiziert das

im Studium erworbene Wissen und die erlernten Fähigkeiten. Aber letztlich spiegeln sich in den Noten auch Aspekte wie Durchhaltevermögen und Selbstorganisation wider, die im Zusammenhang mit nicht-kognitiven Konstrukten (z. B. Selbstwirksamkeit, Gewissenhaftigkeit) stehen (Richardson et al., 2012).

Bei der Auswahl von Studierenden haben vorhergehende Schulleistungen in Form von Noten gute prädiktive Validität bezüglich des Studienerfolgs (Schuler et al., 1990; Trapmann, Hell, Weigand, & Schuler, 2007). Ob sich diese Befunde auf die Auswahl von Absolventen übertragen lassen, wird in dieser Arbeit geprüft.

5.1.2 Persönlichkeitseigenschaften als Leistungsprädiktoren

Sowohl Schul- als auch Studienleistungen beinhalten neben einer kognitiven Leistungskomponente auch motivationale Aspekte (F. Fischer et al., 2012) und Persönlichkeitsmerkmale, die das Lernverhalten und die Arbeitsorganisation beeinflussen (Richardson et al., 2012). Im Fokus dieser Arbeit steht das Faktoren-Modell der Big Five, das fünf überwiegend voneinander unabhängige Persönlichkeitseigenschaften postuliert (Costa & McCrae, 1992). Dabei handelt es sich um Extraversion, Verträglichkeit, Gewissenhaftigkeit, Neurotizismus und Offenheit für Erfahrungen. Diese fünf Persönlichkeitsdimensionen konnten in vielen Untersuchungen empirisch repliziert werden, wobei auch immer wieder der Zusammenhang mit Bildungserfolg untersucht wurde (De Raad & Schouwenburg, 1996). Gewissenhaftigkeit zeigt dabei als einziger Faktor eine klare prädiktive Validität hinsichtlich des Studienerfolgs (Richardson et al., 2012; Trapmann, Hell, Hirn, & Schuler, 2007). Bei der Prognose der beruflichen Zufriedenheit haben emotionale Stabilität und Extraversion konsistent prädiktive Validität (Judge et al., 2002) und zumindest bei Studien mit großen Stichproben gibt es auch einen stabilen Effekt beim Zusammenhang von Gewissenhaftigkeit und Arbeitszufriedenheit (Judge et al., 2002; Lapierre & Hackett, 2007). Ein sehr ähnliches Bild zeigt sich für die Vorhersage von Einkommen: Auch hier finden sich moderat große Korrelationen, die für Gewissenhaftigkeit weniger stabil sind als für emotionale

Stabilität und Extraversion (Judge et al., 1999).

5.1.3 Berufserfolg

Berufserfolg hat viele Gesichter. Man kann grob zwischen intrinsischen (z. B. Zufriedenheit, Passung) und extrinsischen Erfolgskriterien (z. B. Einkommen, Beurteilung durch Vorgesetzte) unterscheiden. Die beiden Dimensionen lassen sich faktorenanalytisch voneinander trennen (Judge et al., 1999).

5.1.4 Geschlechtsunterschiede

In Deutschland wie auch in den meisten anderen europäischen Ländern sind Frauen im Schnitt zufriedener mit ihrer Arbeit als Männer (Kaiser, 2005), allerdings schrumpft der Effekt mit zunehmendem Bildungsgrad (A. E. Clark, 1997) und verschwindet bei Kontrolle der fachlichen Ausrichtung des Berufsfelds offenbar ganz (Abele & Spurk, 2009).

Bei der Betrachtung des Zusammenhangs zwischen den Big Five und dem beruflichen Status zeigen sich mögliche Ursachen für Geschlechtsunterschiede bei der beruflichen Karriere. Das Erreichen einer Führungsposition hängt mit der emotionalen Stabilität, der Offenheit für Erfahrungen, dem Mangel an Verträglichkeit und Gewissenhaftigkeit zusammen, wobei die Effekte in geschlechtsspezifischen Modellen weitgehend verschwinden. Bei den Frauen bleibt einzig der negative Effekt von Verträglichkeit signifikant und bei den Männern bleiben die Effekte von Gewissenhaftigkeit und Verträglichkeit. Emotionale Stabilität und Offenheit für Erfahrungen zeigen bei Kontrolle des Geschlechts keine signifikanten Effekte mehr (Fietze, Holst, & Tobsch, 2010). Entsprechend empfiehlt es sich, bei der Untersuchung von subjektiven Berufserfolgsmaßen den beruflichen Status mit zu berücksichtigen, da Führungspositionen überproportional von Männern besetzt sind.

Geschlechtsspezifische Unterschiede hinsichtlich der beruflichen Position und des Berufsfelds erklären auch einen Teil der sogenannten Lohnlücke. Männer tendieren eher zu Arbeitsgebieten mit höheren Verdienstmöglichkeiten (z.B. Ingenieurberufe), während Frauen eher Berufsfelder mit niedrigeren Einkommen wählen. Berücksichtigt man die entsprechenden Kontroll-

variablen, lässt sich der geschlechtsspezifische Einkommensunterschied um mehr als die Hälfte auf 12 % reduzieren (C. Anger & Schmidt, 2010). Das heißt, dass selbst bei Konstanthaltung der Arbeitsplatz- und der Personeneigenschaften Frauen im Schnitt nur 88 % des Einkommens eines vergleichbaren Mannes erhalten.

5.1.5 Offene Forschungsfragen

Während die eingangs vorgestellten Meta-Analysen einen guten Überblick über die allgemeine Situation geben, fehlen bislang Forschungsarbeiten, die speziell die Erfolgschancen von Hochschulabsolventen untersuchen. Befragungen von Absolventen und rückblickende Interviews mit Berufstätigen können jeweils nur Erkenntnisse über einen zeitlich und räumlich beschränkten Personenkreis liefern (z. B. für eine Hochschule oder für einen Berufszweig). Deshalb wird in der vorliegenden Arbeit anhand von repräsentativen Paneldaten aus Deutschland untersucht, wie gut Schulnoten und Persönlichkeitseigenschaften subjektive (Arbeitszufriedenheit) und objektive (Einkommen) Aspekte der Berufstätigkeit von Hochschulabsolventen zwei Jahre nach deren Abschluss vorhersagen können.

5.1.6 Hypothesen und explorative Annahmen

Da allgemeine kognitive Fähigkeiten (erfasst durch Studierfähigkeitstests und Schulnoten) und nicht-kognitive Facetten (Gewissenhaftigkeit) den Studienerfolg begünstigen (Richardson et al., 2012; Trapmann, Hell, Hirn, & Schuler, 2007; Trapmann, Hell, Weigand, & Schuler, 2007), wird erwartet, dass die Prognosekraft von Schulnoten und den Big-Five-Faktoren in dieser Stichprobe aufgrund der Vorselektion abnimmt. Wenn – stark vereinfacht ausgedrückt – nur Studierende mit einem 1,0-Abitur ein bestimmtes Studium abschließen, kann man anhand der Schulnote nicht mehr zwischen Bewerbern differenzieren. Eine Bestätigung dieser Hypothese würde eine Berücksichtigung von Schulnoten und Persönlichkeitsmaßen bei der Personalauswahl fraglich erscheinen lassen, zumal der Einbezug von Persönlichkeitstests mit zeitlichem und finanziellem Aufwand verbunden ist.

Weiterhin werden geschlechtsspezifische differenzielle Effekte der Schulleistung und der Persönlichkeitsvariablen hinsichtlich der Jobzufriedenheit und des Einkommens getestet. Diese Analyse hat explorativen Charakter, da zu diesen Zusammenhängen bislang keine Studien für Hochschulabsolventen vorliegen.

Unterschiede bei der Wahl des Berufsfelds sind in einer erst kürzlich erhobenen Hochschulabsolventenstichprobe zu erwarten, da generell Frauen stärker zu sozialen Tätigkeiten neigen, während Männer eher zu technischen Arbeiten tendieren (Su, Rounds, & Armstrong, 2009). Um diese Unterschiede bei den beruflichen Interessen zu berücksichtigen, wird die Interessendimension der gewählten Studienrichtung als Kovariate mit einbezogen. Bei den Analysen werden weiterhin beruflicher Status und Abschlussjahr als Kontrollvariablen berücksichtigt.

Bildungsdifferenzen dürften dagegen angesichts der diesbezüglich homogenen Stichprobe vernachlässigbar sein. Persönlichkeitseigenschaften wird ein eher kleiner Einfluss auf die Lohnlücke zugeschrieben, jedoch bleibt besonders Gewissenhaftigkeit ein wichtiger Faktor, da sie das Einkommen möglicherweise indirekt über das Arbeitszeitmanagement beeinflusst (Fietze et al., 2010). Im Mittelpunkt dieser Auswertungen stehen die Forschungsfragen, wie weit Schulleistungen (Noten) und Persönlichkeitseigenschaften (Big Five) mit dem Berufserfolg von Hochschulabsolventen in Deutschland zusammenhängen und welche Geschlechtseffekte es dabei gibt.

5.2 Methode

Die vorliegende Studie analysiert die Daten des Sozio-Ökonomischen Panel (SOEP, Daten für die Jahre 1984–2010, Version 27, 2011, doi:10.5684/soep.v27). Das SOEP liefert Längsschnittdaten basierend auf einer jährlichen, seit 1984 stattfindenden Wiederholungsbefragung von aktuell über 22000 Personen, die zunehmend auch psychologische Themen abdeckt (Schupp, 2009).

Für die Operationalisierung der eingangs vorgestellten Konstrukte muss auf die im Panel in der Vergangenheit erhobenen Variablen zurückgegriffen werden. Diese bieten oftmals nur wenige oder grob differenzierende Items, so

dass die ausgewählten Instrumente nicht immer eine völlig zufriedenstellende Messung, zumindest aber eine gute Annäherung bieten.

5.2.1 Instrumente

Im biographischen Interview, das mit Kindern von Teilnehmern bei der Aufnahme ins Panel im Alter von 17 Jahren geführt wird, werden seit 2001 die jeweils letzten Schulnoten in Mathematik, Deutsch und der ersten Fremdsprache erhoben. Die Noten reichen von 1 („sehr gut“) bis 6 („ungenügend“).

Das Big Five Inventory (BFI-S) erfasst im SOEP die Persönlichkeitseigenschaften mit jeweils drei 7-stufigen Items pro Faktor. Das Instrument erlaubt hinsichtlich der Validität und Reliabilität eine akzeptable Kurzmessung der interessierenden Konstrukte (Hahn, Gottschling, & Spinath, 2012). Die Big Five wurden bislang 2005 und 2009 erhoben. Zur Erstellung der Variablen in dieser Studie wurden die Antworten erst innerhalb der beiden Wellen gemittelt und dann – so denn die betreffende Person in beiden Wellen befragt wurde – über beide Messzeitpunkte hinweg gemittelt. Da die Big Five im Zeitverlauf relativ stabil sind (Donnellan & Lucas, 2008), sind keine nennenswerten Verzerrungen bei den Ergebnissen zu erwarten.

Arbeitszufriedenheit wird jährlich auf einer Skala von 0 (ganz und gar unzufrieden) bis 10 (ganz und gar zufrieden) abgefragt. Dabei wird die berufliche Zufriedenheit im Allgemeinen erfasst („Wie zufrieden sind Sie gegenwärtig mit den folgenden Bereichen Ihres Lebens? – mit Ihrer Arbeit?“). Um eine vergleichbare Einkommensvariable („Stundenlohn“) zu bekommen, wurde das aus den Einzelangaben der Befragten generierte Bruttomonatsgehalt durch 4 und durch die vereinbarte Wochenarbeitszeit geteilt. Die entstandene Stundenlohn-Variable hat wenige Ausreißer nach oben, ist ansonsten aber näherungsweise normalverteilt. Die nachfolgend berichteten Analysen verändern sich nicht wesentlich, wenn die Ausreißer getrimmt werden (d. h. konkret, dass die Gewichte im höchsten 10 %-Perzentil auf den Wert des 90 %-Perzentils herabgesetzt werden).

Beruflicher Status, ein weiteres extrinsisches Berufserfolgskriterium, ist eng verbunden mit Einkommen (Judge et al., 1999). Da die vorliegende Stu-

die Hochschulabsolventen wenige Jahre nach ihrem Abschluss untersucht und die Möglichkeiten zum beruflichen Aufstieg innerhalb dieses Zeitraums beschränkt sind, wird beruflicher Status hier lediglich als Kontrollvariable verwendet. In den SOEP-Daten gibt es als Maß für den beruflichen Status den International Socio-Economic Index of Occupational Status (ISEI; Ganzeboom & Treiman, 1996), der sich aus der Berufsklassifizierung ISCO88 ableitet und von 16 (niedrig) bis 90 (hoch) reicht.

Die SOEP-Angaben zum Studienfeld, in dem der Hochschulabschluss erworben wurde, wurden gemäß dem RIASEC-Interessen-Modell (Holland, 1997) einer der folgenden sechs Interessendimensionen zugeteilt: Realistic (technisch), Investigative (forschend), Artistic (künstlerisch), Social (sozial), Enterprising (unternehmerisch) oder Conventional (traditionell). Die genaue Klassifizierung erfolgte gemäß dem Handbuch des Allgemeinen Interessen-Struktur-Tests – Revision (AIST-R; Bergmann & Eder, 2005).

Während durch die Kontrollvariable Alter mögliche berufliche Vorerfahrungen mit abgedeckt werden, erlaubt das Jahr des Hochschulabschlusses die Berücksichtigung von Inflationseffekten bei Variablen wie Einkommen.

5.2.2 Hochschulabschlussidentifikation und Einschlusskriterien

Aufgrund der oben dargestellten Einschränkungen der Notenvariablen eignet sich der zeitliche Abstand von zwei Jahren zwischen Hochschulabschluss und Berufserfolgserfassung als Kompromiss. Ein kürzerer Abstand kann angesichts des jährlichen Erhebungsplans zu ungewünschten Überlappungen und schlimmstenfalls zu konsistenten Ergebnissen führen. Ein längerer Abstand führt dagegen rasch zu inakzeptabel kleinen Reststichproben.

Da manche Personen zeitlich versetzt einen oder mehrere weitere Abschlüsse erworben haben, wurden zunächst alle Abschlüsse kodiert, ehe mittels der zeitlichen Abstände zwischen den Abschlüssen, der jeweils "letzteÄbschluss einer Person vor dem Verlassen der Hochschule ermittelt wurde. Dafür wurden alle Abschlüsse, die maximal drei Jahre vor dem vorletzten erworben wurden, „gestrichen“ (z. B. Bachelor-Abschlüsse, auf die ein Master-Abschluss folgte). Entsprechend gibt es nach diesem Vorgehen keine Person

mehr, die zwei Jahre nach ihrem „letzten“ Abschluss noch einen Abschluss macht. Wenn der Abstand zwischen zwei Abschlüssen vier Jahre oder mehr beträgt, wird der nachfolgende Abschluss als „Zweitstudium“ interpretiert, welches nicht weiter in die Analyse einfließt.

In die Analysestichprobe wurden alle Personen eingeschlossen, die während der Teilnahme am Panel erfolgreich ihr (im obigen Sinne „letztes“) Studium abschlossen und auch noch zwei Jahre später bei der Befragung teilnahmen. Bei den Angaben, die nicht in allen Wellen erfasst wurden (z. B. Schulnoten, Big Five), treten im Datensatz entsprechend oft fehlende Werte auf. Beobachtungen mit fehlenden Werten werden fallweise von den Analysen ausgeschlossen.

5.2.3 Datenanalyse

Die Vorhersage von Berufserfolg wird mit multiplen Regressionen modelliert. Für die differenziellen Geschlechtseffekte werden moderierte multiple Regressionen (MMRs) gerechnet. Alle kontinuierlichen Prädiktorvariablen werden vorab zentriert, um eine klare Interpretation der Haupteffekte in den MMRs zu erlauben (Aiken & West, 1991).

Die SOEP-Daten gehen größtenteils auf geschichtete Zufallsstichproben aus der in der Bundesrepublik Deutschland lebenden Bevölkerung zurück. Die aus dem Stichprobenplan abgeleiteten Gewichte wurden anhand der Bleibewahrscheinlichkeit fortgeschrieben (Details siehe Kroh, 2010). Da die langfristige Panel-Teilnahme ein Stück weit mit den interessierenden Variablen zusammenhängen, kann es ohne Gewichtung zu systematischen Verzerrungen bezüglich einzelner Personengruppen kommen kann (Schnell, Hill, & Esser, 2005). Deshalb wird bei den Regressionsanalysen hier jede Beobachtung mit ihrem Gewicht zum Berufserfolgsmesszeitpunkt (Inverse der Auswahlwahrscheinlichkeit mal Inverse der Bleibewahrscheinlichkeit bis dahin) gewichtet, um möglichst repräsentative Ergebnisse zu erhalten. Dafür werden „probability weights“ in Stata 12.1 verwendet, die die Gewichtung auch bei der Berechnung der Standardfehler berücksichtigen. Die Gewichte sind im vorliegenden Fall rechtsschief verteilt. Die Trimmung der obersten 10 % führt

jedoch nicht zu gravierenden Veränderungen der Ergebnisse, weshalb die ursprünglichen Gewichte bei den nachfolgenden Regressionen beibehalten wurden.

Da die berufliche Entwicklung ein dynamischer Prozess ist, ergibt sich die Frage, ob – und wenn ja in welcher Form – sich die verschiedenen Aspekte, die den Berufserfolg ausmachen, gegenseitig beeinflussen (L. Fischer & Fischer, 2005). Da die kausalen Zusammenhänge zwischen Arbeitszufriedenheit und Einkommen bislang nur unzureichend erforscht sind, werden in dieser Studie beide Seiten separat betrachtet.

5.3 Resultate

Insgesamt haben in der SOEP-Stichprobe 1153 Personen (46 % Frauen) im Verlauf ihrer Panelteilnahme einen Hochschulabschluss erworben und nahmen zwei Jahre danach noch an der Befragung teil. Im Schnitt waren die Befragten zum Zeitpunkt ihres Studienabschlusses 29.6 Jahre alt (SD: 5.4), wobei es beim Alter wenige Ausreißer nach oben gibt (Median: 28 Jahre).

5.3.1 Deskriptive Statistiken

Ein Großteil der Absolventen war zwei Jahre nach dem Studium berufstätig. Bei den nicht erwerbstätigen und den teilzeitbeschäftigten Personen kamen auf jeden Mann jeweils fast zwei Frauen. Tabelle 5.1 zeigt die Berufsstatushäufigkeiten nach Geschlecht getrennt.

Auch bei der gewählten Studienrichtung zeigen sich Geschlechtsunterschiede (s. Tabelle 5.2). Frauen stellen die Mehrheit in sozial orientierten Studienfeldern, während Männer eher zu praktisch-technischen und forschenden Fächern tendieren.

Die deskriptive Statistik der interessierenden Variablen ist zusammen mit Tests auf geschlechtsspezifische Mittelwertsunterschiede in Tabelle 5.3 aufgelistet. Die Big-Five-Skalen weisen moderat große Geschlechtsunterschiede auf, wobei stets die Frauen höhere Werte als die Männer angaben. Bei den sprachlichen Schulnoten sind die Frauen besser, während es bei der Mathematiknote keinen signifikanten Unterschied gibt. Weiterhin finden sich die erwarteten

Tabelle 5.1: Berufsstatus zwei Jahre nach dem Hochschulabschluss nach Geschlecht getrennt; in Klammern stehen die gewichteten Häufigkeiten.

	Männer	Frauen
Voll berufstätig	532 (553)	358 (334)
Teilzeitbeschäftigung	33 (48)	52 (39)
Ausbildung	4 (2)	5 (9)
Unregelmäßig, gering verdienend	9 (9)	17 (19)
Nicht erwerbstätig	50 (48)	93 (92)
Gesamt	628 (660)	525 (493)

Anmerkung: $\chi^2(4) = 44.93$, $p < .001$ (ungewichtet)

Geschlechtsunterschiede beim beruflichen Status und beim Lohn. Zudem waren Frauen zum Zeitpunkt des Abschlusses im Schnitt jünger und studierten im ersten Jahrzehnt nach Beginn des SOEP noch seltener als Männer.

5.3.2 Bivariate Zusammenhänge

Die ungewichteten Validitätskoeffizienten zur Prognose von Berufserfolg sind in Tabelle 5.4 aufgeführt. Abgesehen von Offenheit für Erfahrung zeigen Big-Five-Faktoren signifikante Zusammenhänge mit der Arbeitszufriedenheit. Bei den Männern hat die Mathematiknote prädiktive Validität für den Berufserfolg. Bei den sprachlichen Noten gibt es dagegen keine klaren Zusammenhänge. Die Kontrollvariablen hängen deutlich mit dem Einkommen zusammen. Die beiden Berufserfolgskriterien Arbeitszufriedenheit und Einkommen korrelieren zudem leicht positiv miteinander.

5.3.3 Prognose von Arbeitszufriedenheit zwei Jahre nach dem Abschluss

Insgesamt werden vier Regressionsmodelle aufgestellt, die in Tabelle 5.5 zusammengefasst sind. Das erste Regressionsmodell zur Vorhersage von Arbeitszufriedenheit verwendet als Prädiktoren die Big-Five-Skalen, Geschlecht,

Tabelle 5.2: Interessendimension des Studienfelds nach Holland (1997) getrennt nach Geschlecht

	Männer	Frauen
Realistic	125	26
Investigative	136	59
Artistic	18	46
Social	60	134
Enterprising	91	70
Conventional	48	47
Gesamt	478	382

Anmerkung: $\chi^2(5) = 129.44$, $p < .001$

die Interessendimension des Studienfelds, die Kontrollvariablen und außerdem als kognitiven Leistungsindikator die Schulnote in Mathematik. Letztere ist wie auch schon im bivariaten Test kein signifikanter Prädiktor und wird deshalb im zweiten Modell weggelassen. Da die verbleibenden Variablen weniger fehlende Werte aufweisen als die Noten, kann für Modell 2 eine wesentlich größere Stichprobe verwendet werden. Dadurch wird der Modellfit allerdings schlechter und lediglich Gewissenhaftigkeit zeigt noch einen signifikanten Effekt. Die Hinzunahme von Interaktionseffekten zwischen den Persönlichkeitsvariablen und Geschlecht in Modell 3 führt zu einem ähnlichen Bild. Lediglich ein positiver Zusammenhang von Verträglichkeit und Arbeitszufriedenheit bei Frauen findet sich hier. Ansonsten gibt es keine Geschlechtsunterschiede. Modell 4 ist eine alternative Erweiterung des zweiten Modells, bei dem das parallel zur Arbeitszufriedenheit erhobene Einkommen als Prädiktor mit eingeschlossen wird. Die signifikante Varianzaufklärung durch das Einkommen illustriert, dass die beiden Berufserfolgskriterien in der untersuchten Stichprobe nicht unabhängig voneinander sind.

Tabelle 5.3: Deskriptive Statistiken für Männer und Frauen; positive d-Werte zeigen höhere Werte für Männer an.

	Männer			Frauen			
Variable	n	M	SD	n	M	SD	Cohen's d
Neurotizismus	485	3.48	1.01	445	3.92	1.09	−.42***
Extraversion	485	4.64	1.14	445	5.07	1.06	−.39***
Offenheit	485	4.59	0.92	445	4.89	1.13	−.29***
Verträglichkeit	485	5.23	0.82	445	5.40	0.82	−.21**
Gewissenhaftigkeit	485	5.65	0.80	445	5.84	0.79	−.23***
Mathematiknote	155	2.22	1.03	166	2.43	1.05	−.21
Deutschnote	150	2.49	0.83	166	2.08	0.88	.47***
Fremdsprachennote	153	2.53	0.93	166	2.17	0.89	.39***
Beruflicher Status (ISEI)	563	64.1	13.2	421	62.3	14.2	.13*
Arbeitszufriedenheit	582	7.25	1.83	444	7.20	1.84	.03
Einkommen (in Euro)	465	19.1	7.46	357	16.4	6.04	.38***
Alter bei Abschluss	628	30.1	5.45	525	28.9	5.21	.22***
Abschlussjahr	628	1998	6.41	525	2000	6.35	−.22***

Anmerkungen. Mittelwertvergleich mittels T-Test:
* $p < .05$, ** $p < .01$, *** $p < .001$

Tabelle 5.4: Validitätskoeffizienten bei der Vorhersage von Berufserfolg

	Männer				Frauen			
Variable	n	AZ	n	Eink.	n	AZ	n	Eink.
Neurotizismus	457	−.18***	371	−.12*	384	−.18***	303	.01
Extraversion	457	.13**	371	.04	384	.15**	303	.01
Offenheit	457	.08	371	.04	384	.05	303	.04
Verträglichkeit	457	.10*	371	−.06	384	.23***	303	−.05
Gewissenhaftigkeit	457	.20***	371	.07	384	.16**	303	−.07
Mathematiknote	149	−.24**	114	−.25**	147	−.08	111	−.11
Deutschnote	144	.01	111	−.07	147	−.15	112	−.07
Fremdsprachennote	148	−.13	113	−.00	147	−.04	111	−.03
Beruflicher Status (ISEI)	560	.08	456	.23	415	.10	350	.33***
Arbeitszufriedenheit			462	.10*			353	.16**
Einkommen (in Euro)	462	.10*			353	.16**		
Alter bei Abschluss	582	−.08	465	.16***	444	.04	357	.34***
Abschlussjahr	582	−.03	465	.31***	444	.07	357	.30***

Anmerkungen. AZ = Arbeitszufriedenheit, Eink. = Einkommen;
* $p < .05$, ** $p < .01$, *** $p < .001$

Tabelle 5.5: Multiple Regressionsmodelle zur Vorhersage von Arbeitszufriedenheit

Prädiktorvariablen	(1)	(2)	(3)	(4)
Neurotizismus	−0.113	−0.150	−0.050	−0.002
	(0.155)	(0.104)	(0.146)	(0.128)
Extraversion	0.400**	0.182	0.240	0.134
	(0.196)	(0.130)	(0.192)	(0.141)
Offenheit	0.024	−0.035	−0.066	−0.054
	(0.162)	(0.110)	(0.141)	(0.124)
Verträglichkeit	0.489***	0.218	0.0235	0.238
	(0.187)	(0.133)	(0.191)	(0.145)
Gewissenhaftigkeit	0.580**	0.353**	0.397**	0.389**
	(0.229)	(0.137)	(0.191)	(0.166)
Geschlecht (weiblich)	0.109	−0.009	0.038	0.244
	(0.383)	(0.226)	(0.213)	(0.253)
weiblich*Neurotizismus			−0.280	
			(0.203)	
weiblich*Extraversion			−0.155	
			(0.233)	
weiblich*Offenheit			0.0809	
			(0.210)	
weiblich*Verträglichkeit			0.431*	
			(0.259)	
weiblich*Gewissenhaftigkeit			−0.129	
			(0.263)	
Beruflicher Status	−0.002	0.009	0.011	−0.004
	(0.021)	(0.010)	(0.010)	(0.012)
Investigative-Dummy	0.168	−0.023	−0.116	−0.083
	(0.660)	(0.339)	(0.327)	(0.356)
Artistic-Dummy	0.451	−0.115	−0.148	−0.188
	(0.673)	(0.402)	(0.397)	(0.539)

(Fortsetzung Tabelle 5.5)

Prädiktorvariablen	(1)	(2)	(3)	(4)
Social-Dummy	−0.437	−0.218	−0.267	−0.0729
	(0.730)	(0.409)	(0.400)	(0.394)
Enterprising-Dummy	0.568	−0.306	−0.360	−0.482
	(0.580)	(0.327)	(0.319)	(0.340)
Conventional-Dummy	0.255	0.156	0.149	0.447
	(0.547)	(0.334)	(0.327)	(0.397)
Alter beim Abschluss	−0.043	−0.018	−0.019	−0.062**
	(0.041)	(0.021)	(0.021)	(0.027)
Abschlussjahr	0.047	0.001	0.003	−0.013
	(0.064)	(0.016)	(0.016)	(0.022)
Mathematiknote	−0.194			
	(0.220)			
Einkommen				0.063**
				(0.028)
Konstante	6.765***	7.373***	7.421***	7.187***
	(0.587)	(0.226)	(0.227)	(0.228)
Beobachtungen	194	589	589	481
R^2	0.281	0.094	0.110	0.106

Anmerkungen. Standardfehler in Klammern;
* $p < .05$, ** $p < .01$, *** $p < .001$

5.3.4 Prognose des Einkommens zwei Jahre nach dem Abschluss

Für die Vorhersage des Einkommens nach dem Studium werden ebenfalls vier verschiedene Regressionsmodelle gerechnet (s. Tabelle 5.6). Zuerst werden nur die Schulleistung, die Interessenorientierung im Studium sowie die Kontrollvariablen eingeschlossen (Modell 1). Die Lohnlücke zwischen Frauen und Männern ist klar vorhanden. Ebenso gibt es Einkommensunterschie-

de zwischen verschiedenen Interessendimensionen, wobei die teilweise sehr großen Regressionskoeffizienten auch teilweise auf niedrige Zellbesetzungen zurückzuführen sind; so gibt es beispielsweise keinen Mann in der Teilstichprobe, der Artistic-orientiert studiert hat. Während die Mathematiknote bei Kontrolle der anderen Variablen signifikante prädiktive Validität zeigt, sind die Effekte der Noten in den Sprachenfächern kleiner und weniger eindeutig. Deshalb wird in Modell 2 nur die Note in Mathematik berücksichtigt. Zusätzlich werden die Big-Five-Skalen mit aufgenommen. Es zeigt sich ein signifikanter positiver Zusammenhang zwischen Verträglichkeit und Einkommen sowie ein negativer Zusammenhang zwischen Gewissenhaftigkeit und Gehalt. Die Interaktionsterme von Geschlecht und Persönlichkeit, die in Modell 3 getestet werden, relativieren diese Haupteffekte jedoch. Der Effekt von Verträglichkeit geht offenbar auf die Frauen zurück. Da die ersten drei Modelle wegen der spärlichen Noteninformationen nur auf einer kleinen Teilstichprobe basieren, enthält Modell 4 nur die Persönlichkeitsvariablen, die Interessenrichtung des Studienfelds sowie die Kontrollvariablen als Prädiktoren. Trotz der größeren Fallzahl werden die Koeffizienten der Big-Five-Skalen nicht signifikant. Zudem fallen die Unterschiede zwischen den Interessendimensionen und den Geschlechtern kleiner aus.

Tabelle 5.6: Multiple Regressionsmodelle zur Vorhersage von Einkommen

Prädiktorvariablen	(1)	(2)	(3)	(4)
Neurotizismus		−0.217	−0.277	−0.519
		(0.617)	(1.243)	(0.455)
Extraversion		1.183	1.761	0.377
		(0.785)	(1.232)	(0.456)
Offenheit		0.016	−0.159	−0.070
		(0.582)	(1.110)	(0.497)
Verträglichkeit		1.545**	0.078	−0.355
		(0.676)	(1.262)	(0.682)
Gewissenhaftigkeit		−1.784**	−0.875	−0.388
		(0.714)	(1.407)	(0.452)
Geschlecht (weiblich)	−4.337**	−4.776***	−4.313***	−2.609***
	(1.674)	(1.387)	(1.400)	(0.970)
weiblich*Neurotizismus			0.351	
			(1.358)	
weiblich*Extraversion			−1.966	
			(1.246)	
weiblich*Offenheit			0.192	
			(1.333)	
weiblich*Verträglichkeit			2.666*	
			(1.503)	
weiblich*Gewissenhaftigkeit			−1.722	
			(1.607)	
Mathematiknote	−2.075**	−2.747***	−2.276**	
	(0.890)	(0.933)	(0.875)	
Deutschnote	0.493			
	(1.123)			
Fremdsprachennote	−1.138			
	(0.874)			

(Fortsetzung Tabelle 5.6)

Prädiktorvariablen	(1)	(2)	(3)	(4)
Beruflicher Status	0.158***	0.148**	0.177***	0.182***
	(0.055)	(0.063)	(0.060)	(0.034)
Investigative-Dummy	4.164	2.975	2.298	2.369
	(2.569)	(1.996)	(1.912)	(1.484)
Artistic-Dummy	8.824***	8.058***	5.951**	3.722**
	(3.119)	(2.466)	(2.537)	(1.549)
Social-Dummy	2.840	1.620	0.982	0.308
	(2.538)	(2.187)	(2.156)	(1.541)
Enterprising-Dummy	8.248***	7.478***	5.847***	2.795*
	(2.688)	(2.252)	(2.145)	(1.508)
Conventional-Dummy	3.484	1.725	0.050	−1.678
	(2.985)	(2.528)	(2.460)	(1.498)
Alter beim Abschluss	0.395**	0.441***	0.436***	0.360***
	(0.163)	(0.167)	(0.155)	(0.093)
Abschlussjahr	0.218	0.077	0.213	0.277***
	(0.260)	(0.231)	(0.248)	(0.061)
Konstante	16.12***	18.56***	18.32***	18.18***
	(2.023)	(1.666)	(1.832)	(0.981)
Beobachtungen	157	154	154	484
R^2	0.371	0.394	0.435	0.336

Anmerkungen. Standardfehler in Klammern;
* $p < .05$, ** $p < .01$, *** $p < .001$

5.4 Diskussion

Mit Hilfe von Paneldaten wurde in dieser Studie untersucht, ob typische Berufserfolgsprädiktoren bei Personen mit Hochschulabschluss noch prädiktive Validität besitzen und wie weit geschlechtsspezifische Unterschiede auftre-

ten. Für die Vorhersage von Arbeitszufriedenheit durch Persönlichkeitseigenschaften eignet sich Gewissenhaftigkeit und bei Frauen zusätzlich Verträglichkeit. Auch bei der Vorhersage des Einkommens ist Verträglichkeit bei Frauen ein valider Prädiktor. Die Validität der Mathematiknote wird durch die Vorselektion im Studium offenbar kaum geschmälert und ist vergleichbar mit meta-analytischen Befunden aus breiteren Populationen (Roth & Clarke, 1998). Die Diskrepanz zwischen dem Durchschnittsgehalt von Männern und dem von Frauen deckt sich ebenfalls mit Befunden aus der gesamten Arbeitsbevölkerung (Fietze et al., 2010). Auch in den hier aufgestellten Modellen kann die Lohnlücke durch Interessenunterschiede und beruflichen Status nicht vollständig aufgeklärt werden.

5.4.1 Berufserfolgsprognose

Der erfolgreiche Abschluss eines Hochschulstudiums ist ein wichtiges Selektionskriterium bei der Berufswahl. Zum einen werden erworbene Fähigkeiten und erlerntes Wissen für die Ausübung anspruchsvoller Berufe benötigt, zum anderen ist der Abschluss auch ein Indikator für eine gute Person-Umwelt-Passung von persönlichen Interessen und Studiums- sowie Arbeitsinhalten, die sich positiv auf die spätere Arbeitszufriedenheit auswirkt (Wolniak & Pascarella, 2005). Durch diese (Selbst-)Selektion reduziert sich die Varianz vieler Merkmale in der Absolventenpopulation im Vergleich zur Gesamtbevölkerung. Die bestenfalls schwachen Effekte von Extraversion in den Regressionen deuten darauf hin, dass Befunde zur allgemeinen Berufserfolgsprognose mittels Big-Five-Skalen (z. B. Barrick, Mount, & Judge, 2001) nicht blind auf die Auswahl von akademisch gebildetem Personal übertragen werden sollten. Innerhalb der Gruppe der Studienabgänger gibt es keine nennenswerten Geschlechtsunterschiede bei der Arbeitszufriedenheit, was sich mit Ergebnissen aus vergleichbar homogenen Stichproben deckt (Abele & Spurk, 2009).

Die Analyse der Daten von Hochschulabsolventen im SOEP zeigt zugleich Möglichkeiten auf, wie Leistungs- und Persönlichkeitsvariablen gezielt zur Verbesserung der Vorhersagemodelle verwendet werden können. Der intrinsi-

sche Berufserfolg in Form von Arbeitszufriedenheit hängt mit Gewissenhaftigkeit zusammen, wobei zu beachten ist, dass die drei Items der BFI-S-Skala (z. B. „Ich bin jemand, der Aufgaben wirksam und effizient erledigt.“) in erster Linie Selbstdisziplin und Pflichtbewusstsein erfassen (Hahn et al., 2012). Die Validitätsbefunde können nicht auf andere Unterfacetten der Gewissenhaftigkeit wie beispielsweise Ordnungsdrang generalisiert werden.

Während die Arbeitszufriedenheit über verschiedene Studienfelder hinweg im Schnitt ähnlich hoch ist, zeigen sich beim Einkommen deutliche Unterschiede. Durch die verschiedenen fachlichen Interessen von Männern und Frauen schleicht sich hier auch ein Geschlechtseffekt ein. Hinzu kommt, dass der berufliche Status in den Bereichen Artistic und Enterprising im Schnitt niedriger ist als in der Referenzgruppe Realistic (allerdings nicht signifikant). Bei Konstanthaltung der Kontrollvariablen ist die Schulnote in Mathematik ein valider Prädiktor des Einkommens zwei Jahre nach dem Studium. Da Mathematiknoten in hohem Maß allgemeine Intelligenz abbilden, deckt sich dieses Ergebnis mit den ersten Befunden der Leistungskurztests im SOEP (Teilstichprobe Westdeutschland, 20–60-Jährige, Erhebungsjahr 2005), die zeigen, dass kognitive Fähigkeiten einen Einfluss auf das Einkommen haben, der über die Bildungsrendite hinausgeht (S. Anger & Heineck, 2010). Bei der Auswahl von Hochschulabsolventen ist ein gezielter Blick ins Abiturzeugnis der Bewerber also durchaus sinnvoll.

Weiterhin zeigen sich beim Einkommen deutliche Geschlechtsunterschiede zu Gunsten der Männer. Die Lohnlücke beträgt in der Referenzgruppe (Realistic) bei durchschnittlicher Ausprägung aller anderen Variablen etwa 14 %. Eine studierte Frau erhält im Schnitt also nur 86 % des Einkommens eines vergleichbaren Mannes. Das entspricht näherungsweise dem Gehaltsunterschied, der nach Berücksichtigung von Berufsstatus, Ausbildung und Fachgebiet in Studien der berufstätigen Bevölkerung noch bleibt (12.9 % bei Anger & Schmidt, 2010; 12 % bei Hinz & Gartner, 2012). Berücksichtigt man Studienfeld-Geschlecht-Interaktionen im Regressionsmodell, ergeben sich hochgerechnet für das Jahr 2008 (dem letzten in dieser Studie berücksichtigten Abschlussjahr) für die einzelnen Interessendimensionen folgende Einkommensunterschiede: Realistic 29.8 %, Investigative −1.5 %, Artistic 12.7 %, Social

14.3 %, Enterprising 10.9 % und Conventional 14.0 %. Angesichts der ungleichen Fachpräferenzen von Männern und Frauen (s. Tabelle 5.2) sind diese Befunde mit Vorsicht zu genießen. Zudem gibt es keine signifikanten Interaktionen ($p = .073$ für den Lückenunterschied zwischen Realistic und Investigative) und das Regressionsmodell weist – im Gegensatz zu den oben berichteten – Multikollinearitätstendenzen auf (Varianzinflations-Faktor > 8 für den Geschlechtseffekt, d. h. der dazugehörige Standardfehler kann durch den relativ großen Zusammenhang von Geschlecht mit den anderen Prädiktorvariablen verzerrt werden; vgl. Eid, Gollwitzer & Schmitt, 2010, S. 686–687).

Mit Bezug auf geschlechtsspezifische Stereotype von Führungspersonen (Fietze et al., 2010) ist der positive Interaktionsterm für Frauen und Verträglichkeit vermutlich schlüssiger dahingehend interpretierbar, dass bei Frauen (jedoch nicht bei Männern) ein Mangel an Verträglichkeit mit Einkommenseinbußen verbunden ist.

5.4.2 Limitationen

Die Operationalisierung der Konstrukte erfolgte anhand weniger Indikatoren, die sicherlich nicht die gesamte Breite der interessierenden Merkmale abdecken, zumal sie nicht speziell für die hier behandelten Forschungsfragen entwickelt wurden. Unterfacetten der Persönlichkeitseigenschaften konnten beispielsweise nicht erhoben werden und die fünf allgemeinen Persönlichkeitseigenschaften decken jeweils nur bestimmte Aspekte ab (Hahn et al., 2012). Faktoren, die in dieser Analyse keine Effekte zeigten, sollten folglich nicht voreilig verworfen werden. Möglicherweise ergeben Untersuchungen mit umfangreicheren Instrumenten ein wesentlich differenzierteres Bild.

Gleichzeitig darf die Bedeutung der gefundenen Zusammenhänge nicht überinterpretiert werden. Es gibt neben Mathematik und Sprachen in der Schule viele andere Fächer. Auch haben andere, hier nicht erfasste psychologische Parameter wie beispielsweise Selbstwirksamkeitserwartungen prognostische Validität hinsichtlich des Berufserfolgs von Hochschulabsolventen (Abele-Brehm & Stief, 2004).

5.4.3 Fazit

Die Analysen des SOEP in diesem Artikel zeigen, dass bei der Rekrutierung von hochgebildetem Nachwuchspersonal Schulleistungen in Mathematik sowie Gewissenhaftigkeit valide Prädiktoren für Berufserfolgskriterien sein können. Aufgrund der Repräsentativität der analysierten Paneldaten und der Abdeckung von zwei wichtigen Berufserfolgskriterien, die hoch mit anderen arbeitspsychologischen Aspekten (z. B. Fremdbeurteilung durch Vorgesetzte) korrelieren (Judge et al., 1999), helfen diese Befunde bei der Gestaltung von Beratung und Personalauswahl von Berufseinsteigern. Außerdem konnte gezeigt werden, dass das Einkommen von Hochschulabgängern auch vom fachlichen Interesse im Studium abhängt.

6 Study 4: Sex-Specific Differential Prediction of College Admission Tests: A Meta-Analysis

Abstract

This is the first meta-analysis that investigates the differential prediction of undergraduate and graduate college admission tests for women and men. Findings on 130 independent samples representing 493,048 students are summarized. The underprediction of women's academic performance ($d = .14$) and the overprediction of men's academic performance ($d = -.16$) are generalizable, albeit small. Transferred onto a four-point grading scale, women earn college grades that are .24 points higher than those of men with the same admission test result. Combining admission tests with indicators of previous academic achievements, such as high school grades, reduces the amount of under- and overprediction. Moderator analysis reveals that the underprediction of women's academic performance by admission tests is a problem of the past and present. Predictor differences (indicating a bias in the test) as well as criterion differences (indicating a bias in the criterion) are not associated with over- and underprediction. Rather, undergraduate college admission tests show more underprediction of women's academic performance than graduate admission tests. These results point to differences between undergraduate and graduate students, the latter being more selected.

[17]This research was supported by the Federal Ministry of Education and Research (grant agreement number: 01FP0930) and the European Social Fund of the European Union.

We thank Sabrina Strohmeier for her help coding articles and Lea Ludwig for her comments on an earlier version of the article.

Keywords: differential prediction, test bias, gender, sex differences, meta-analysis

6.1 Introduction

Every year, millions of people take standardized admission tests in order to be accepted into a college or university. The significant influence of the tests on this key aspect of society has induced a vast amount of research regarding the predictive power of admission tests. The raw correlation between the SAT and first-year college grade point averages (GPA) is .35 and increases to .53 after correction for range restriction[18] (Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008). The Graduate Management Admission Test (GMAT; Kuncel et al., 2007), the Graduate Record Examination (GRE; Kuncel et al., 2010), and subject-specific admission tests in German speaking countries (Hell et al., 2007) show similar results. Although predictive validity is necessary for high stakes testing (see Standards for Educational and Psychological Testing, American Educational Research Association [AERA], American Psychological Association [APA] and National Council on Measurement in Education [NCME], 1999), it is not sufficient for its fairness.

A review of existing literature supports the conclusion that professionally constructed tests are not systematically biased against minority group members in the prediction of academic performance (Linn, 1973; Sackett et al., 2008; Young & Kobrin, 2001). There is, however, evidence that achievement test scores underpredict women's academic performance (Holden, 1989; Young & Kobrin, 2001). In other words, females with the same test scores as males earn better college grades on average. Efforts to summarize the literature in this field are more than ten years old and restricted with regard to content and method. The present study overcomes these limitations by

[18] Analyzing only admitted and enrolled students underestimates the true correlation, since admitted students tend to have a narrower range of test scores than the applicant pool. This problem can be addressed by correcting the correlation for range restriction. The Pearson-Lawley multivariate correction can be applied for this purpose (e. g., Gulliksen, 1950).

providing an up-to-date meta-analysis. International research results with regard to both undergraduate and graduate college admissions are considered and for the first time, group-specific residuals from large-scale studies are summarized with the help of meta-analytic techniques.

6.1.1 Test Fairness and Test Bias in Predicting Subgroups

Regarding the definition of *test fairness* and *test bias* there has been some disagreement in the past. Today, there is consensus that bias relates more to statistical approaches, whereas fairness is a more value-laden concept (Meade & Fetzer, 2009). In the present study, we focus on a bias, which can emerge in the prediction of a subgroup's criterion as defined by Cleary (1968):

> A test is biased for members of a subgroup of the population if in the prediction of a criterion, for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. (p. 115)

This approach is endorsed by the Principles for the Validation and Use of Personnel Selection Procedures (Society for Industrial and Organizational Psychology [SIOP], 2003) and the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999). The Standards conclude that "no bias exists if the regression equations relating the test and the criterion are indistinguishable for the groups in question" (AERA, APA, & NCME, 1999, p. 79).

Although Cleary (1968) and the Standards (AERA, APA, & NCME, 1999) use the term *test bias* "the term *differential prediction* much more accurately describes what is assessed by the regression-based procedure for evaluating the across groups equality of the relationship between the test and the criterion" (Meade & Fetzer, 2009, p. 740). The present study follows this suggestion and applies the term *differential prediction*.

6.1.2 Differences between Differential Prediction and Differential Validity

It is important to distinguish between differential validity and differential prediction because they are obviously related but not identical concepts (Young & Kobrin, 2001). *Differential validity* determines whether the correlations between test results and a criterion are equal across various groups. In contrast, *differential prediction* refers to group differences in regression equations or in standard errors of estimates. Consequently, "equal correlations do not necessarily imply equal standard errors of estimate, nor do they necessarily imply equal slopes or intercepts" (Linn, 1978, p. 511). A test may predict the criterion with the same accuracy for different subgroups but may still underpredict one of these groups.

In empirical test evaluations, the employment of differential validity studies is much more widespread than the employment of differential prediction studies. Differences in prediction, however, have a more direct bearing on considerations of selection (Linn, 1982). The underpredicted group is particularly worrisome, because group members with low scores on the test may not be admitted even though they would perform well at college or university (Huff, Koenig, Treptau, & Sireci, 1999).

6.1.3 How to Measure Differential Prediction

Analyzing differences in regression equations Gulliksen and Wilks (1950) recommended computing separate regression lines for each group and analyzing the components of these regression models sequentially in three steps: (1) compare the standard errors of estimate; (2) test the slope differences, assuming that the errors are equal; and (3) test the intercept differences, assuming that the errors and the slope differences are equal. Since then, this procedure has often been used without step one, testing for differences in the errors of estimate (e. g., Cleary, 1968; Bridgeman & Wendler, 1991). In order to predict academic performance with college admission tests, Cleary implemented this procedure in 1968. From that time on, most of the corresponding studies have referred to Cleary (1968) and have called

this statistical procedure the Cleary approach (e. g., Linn, 1973; Meade & Tonidandel, 2010). Performing a moderated multiple regression and testing its components is equivalent to this procedure (Bartlett et al., 1978). The corresponding formula is

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2, \tag{6.1}$$

in which X_1 is the predictor (e. g., admission test score), X_2 is the group dummy variable (e. g., sex), and $\hat{Y}$ is the predicted criterion (e. g., predicted academic achievement). The group regression lines are regarded as being identical if there are no significant differences between the intercepts and/or slopes of the regression lines for the groups in question. If the intercepts differ but the slopes do not, one can draw a clear conclusion regarding over- and underprediction. The group with the larger intercept is underpredicted, meaning that members of this group perform better than predicted by the test (on average). The group with the smaller intercept is overpredicted by the test. These group members perform worse than predicted. If the slopes differ, regression lines may cross and therefore conclusions regarding over- and underprediction may vary between different test score sections.

Analyzing differences in group-specific residuals As an alternative to the Cleary approach, Lawshe (1983) introduced a simplified procedure. In this method, the mean residuals for every group are calculated based on a common regression line. Negative group residuals indicate overprediction. Positive group residuals indicate underprediction. In some exceptions, the prediction error is calculated by subtracting the actual from the predicted criterion or the algebraic sign of the residuals is changed intentionally in order to align the sign to the meaning (e. g., Bridgeman et al., 2000; M. J. Clark & Grandy, 1984; Talento-Miller, 2008). In large-scale studies, group-specific residuals are preferred to the significance tests of Cleary because statistical tests are prone to indicate significance due to the large test power (e. g., Cohen, 1988).

6.1.4 Previous Efforts to Summarize Sex-Specific Differential Prediction of Admission Tests

In this section, we provide a short summary of two previous reviews that explore differential prediction of admission tests by gender and we show the restrictions of these reviews.

Sanber and Millman (1987) summarized differential prediction effects associated with standardized achievement tests. The authors aggregated 38 studies including 147 samples in an unpublished meta-analysis. Results showed a significant mean slope difference ($M = -.80$; $SD = 1.87$) and no significant intercept difference ($M = .20$; $SD = 2.31$). The method used to compare and aggregate the b values is questionable because there is no statistical justification for a simple summary of slope differences (Becker & Wu, 2007). Since Sanber and Millman (1987) found slope differences, the aggregated results allowed no clear conclusion about over- and underprediction. Therefore, they provided a descriptive summary. Of the summarized samples 81 % reported equal slopes for males and females. In 53 % of these cases, the intercepts were higher for females than for males, indicating marginal underprediction of women and marginal overprediction of men. Still, the results did not allow conclusions about college admission tests in particular, only about achievement tests in general.

In 2001, Young and Kobrin published an extensive summary about the literature on differential prediction in American college admission. The review arrived at the conclusion that the majority of studies reported underprediction of females. More precisely, the mean underprediction of women was about .06 grade points (based on a 0–4 scale). One limitation of this study was that these results were inferred without applying meta-analytic techniques. Therefore, despite the broad scope of the review, conclusions about the generalizability of the results cannot be drawn. Further, results based on different methods that are not necessarily comparable were included. For example, residuals from male regression lines as well as structural equation modeling parameters were summarized and studies using a combination of test scores and high school grades as predictors were not considered sepa-

rately.

6.1.5 The Present Study

There are two major goals in the present study. The first is to examine the general extent of the potential underprediction of women's academic performance and the potential overprediction of men's academic performance by undergraduate and graduate admission tests. As a related question, we investigate whether the combination of high school GPA (HGPA) and undergraduate admission tests, or undergraduate GPA (UGPA) and graduate admission tests reduces the magnitude of differential prediction. Previous research (e. g., Mattern et al., 2008) suggests that combining grades and test scores yields less bias than test scores alone.

Unlike the two studies summarized in the previous section, we focus on undergraduate and graduate tests and we separately investigate predictions based on tests and tests combined with grades. We implement an international perspective by searching established literature data bases that list primary studies from all over the world, like Web of Science and PSYNDEX. We also include recent findings, since many large-scale studies have been published in the last decade. Last but not least, we apply meta-analytic methods to test the generalizability of the results.

If we find differential prediction, the second goal is to improve our understanding regarding the factors that are related to the underprediction of women's performance by identifying potential moderators. Various moderators are discussed in the literature (e. g., Sackett et al., 2008; Zwick, 2002). In the present meta-analysis we focus on the most promising.

Firstly, we look at publication and sample properties. Nowadays items are well reviewed to avoid test content that is more familiar to men or women (Zwick, 2002, p. 152; Educational Testing Service, 2009). To investigate whether this fact has reduced differential prediction changes over time are examined. The mean age of the prospective students is assessed to control for possible gender differences in cognitive development (e. g., Ellis et al., 2008, p. 287; Lynn & Kanazawa, 2011; Lynn & Irwing, 2004).

Secondly, we look at test and grading properties. The test type/name is an obvious candidate regarding moderators because test content plays a crucial role in test outcomes (Zwick, 2002). Similarly, different test components such as verbal and mathematic sections might be linked to differential prediction (e. g., Bridgeman, Pollack, & Burton, 2008; B. F. Patterson et al., 2009).

We also test if differential prediction is related to test score differences between men and women and differences in college grades. According to Meade and Fetzer (2009), if a biased test is responsible for different group regression lines, there is most likely a test score difference (but no relating criterion difference). Conversely, criterion differences (but no relating test score difference) could indicate some type of bias in the criterion.

There is also the possibility that the prediction of final grades is less biased than the prediction of the first year GPA. This could be due to differential dropout rates or due to changing requirements of the learning content (basics vs. special themes). We therefore also test the relation of differential prediction and the time span between predictor and criterion assessment.

Finally, we look at course taking patterns. It was argued that females tend to enroll in less stringent courses with more lenient grading systems (Conger & Long, 2010; Alon & Gelbgiser, 2011). Correcting for differences in grading standards or course taking patterns reduced underprediction of women (Bridgeman et al., 2000; Ramist, C. Lewis, & McCamley-Jenkins, 1994; Willingham, Pollack, & Lewis, 2002). Leonard and Jiang (1999), nevertheless, showed that underprediction of women's grades persists after controlling for gender differences in fields of study and for sample selection bias.

6.2 Method

6.2.1 Literature Search

We used three search strategies to locate published and unpublished studies: (a) database searches of PsycINFO, ERIC, PubMed, PsycARTICLES, Web of Science, PSYNDEX, and Google Scholar using the search terms: (sex or gender) paired with (differential predict* or academic predict* or predict*

bias etc.) paired with (admission test* or placement test* etc.); (b) manual searches through the reference lists of key articles; and (c) screenings of test-homepages and homepages of test providers (e. g., The College Board). The search was conducted at the beginning of 2010. All studies published before then were considered.

6.2.2 Inclusion Criteria

Each of the potential articles was evaluated for inclusion based on the following criteria. Firstly, the study had to examine the prediction of men's and women's college performance by an admission test or a combination of admission test result and previous grades. Secondly, the authors had to report differential prediction results for men and women by: (a) estimating separate regression lines for each gender and comparing their slopes and intercepts (this also includes moderated multiple regression studies with interaction terms); or (b) estimating a joint regression line, analyzing the mean residual for each gender and reporting enough information to calculate effect sizes; or (c) providing all required information to calculate the standardized mean residuals post hoc. Thirdly, the study must have been published in English or in German. We also considered studies written in German, our native language, to further extend the number of potential samples. If the same sample was analyzed in multiple studies, we only included the study that contained most of the relevant data to avoid a duplicate study effect (Wood, 2008).

6.2.3 Summary of the Data Set

The literature search identified 962 studies. Out of this pool, 42 studies met all of the inclusion criteria. The remaining studies could not be included, mainly because they only reported differential validities without providing statistics on differential prediction. Further reasons for exclusion were limited criteria information (e. g., dichotomous pass/fail) and insufficient information about the required statistics for each gender. Also, prediction models that contained additional predictor variables (e. g., personality traits) could not

be statistically disentangled.

The selected studies were published between 1973 and 2009 and they contained 130 samples with a total of 493,048 participants. Group-specific residuals or the information required to calculate them were reported in 28 studies (83 samples). Out of these samples, 55 reported residuals based on an admission test and 52 offered residuals based on a combination of admission test and HGPA/UGPA. Differences in regression equations were reported in 14 studies (47 samples). Apparently, there was an overlap between the studies/samples reporting residuals based on admission tests and studies reporting residuals based on the combination of test scores and HGPA/UGPA. We handled this dependency by separately aggregating the residuals. There was no overlap between samples providing residuals and samples providing differences in regression equations. The criterion was typically first year GPA (more detailed characteristics of the studies, such as author, sample size, and name of the admission test are presented in Tables and A.2).

6.2.4 Coding of Study Variables

Data of independent samples were coded separately. For some samples, all required information was obtained for different predictors and/or criteria. In these cases, the following decision rules were applied. Firstly, the whole test was used as the predictor instead of test parts. Secondly, the criterion with the biggest sample size was chosen. In ambiguous cases, the first year GPA was analyzed instead of later earned grades. Only three studies offered different criteria. We therefore could not make use of multivariate techniques to handle multiple outcomes (e. g., Becker, 2000).

Following aspects were coded as potential moderators: publication year, age of the participants, test type, verbal and mathematic test components, gender differences in test scores, gender differences in HGPA/UGPA, average time span between predictor assessment and criterion assessment, and freedom of course choice. Test score and grade differences were expressed in effect sizes before the moderator analyses were performed. If both statistics were given, we corrected the predictor differences with the help of the

criterion differences and vice versa.

The first and the second author coded all of the studies independently. Both were familiar with the field of study and had created the coding scheme. The initial inter-rater agreement was 96 %. Discrepancies between the raters were solved by consulting a third rater and having discussions to reach a consensus. There were no coded variables with a disproportionate amount of initial differences.

6.2.5 Analytical Procedures

Summarizing residuals In order to aggregate residuals, they have to be transferred to summable statistics. Lawshe (1983) proposed to test whether or not group-specific mean residuals ($\bar{E}_{\text{men}}$ and $\bar{E}_{\text{women}}$) differ with

$$t = [(\bar{E}_{\text{men}} - \bar{E}_{\text{women}})/SD] \cdot \sqrt{N}. \tag{6.2}$$

Unfortunately, the two mean residuals are not independent of each other because

$$N_{\text{men}} \cdot \bar{E}_{\text{men}} + N_{\text{women}} \cdot \bar{E}_{\text{women}} = 0. \tag{6.3}$$

Thus, the assumption of the t-test for independent subgroups is violated. To avoid this problem we do not agree with the proposal from Lawshe (1983). Instead, we recommend the null hypothesis that suggests that the sex-specific residuals do not differ from zero:

$$t_j = [(\bar{E}_j - 0)/SD] \cdot \sqrt{N_j} \qquad (j = \text{men, women}). \tag{6.4}$$

Previous research suggests an underprediction of women's performance, but we do not want to exclude the option that there is indeed an overprediction. The same applies vice versa to men. Therefore, we recommend looking at two-tailed tests.

In the present meta-analysis we did not perform the t-tests but rather calculated the corresponding effect sizes within each sample (for each gender separately). According to Cohen (1988, pp. 45–48) the effect sizes were calculated by

$$d_j = (\bar{E}_j - 0)/SD \qquad (j = \text{men, women}). \tag{6.5}$$

We used the standard deviation of the total sample since the residuals are based on one regression line (this procedure is equivalent to standardizing the residuals). If the total standard deviation was not reported, we computed the pooled standard deviation. In cases where there was no standard deviation available, it was calculated by

$$SD = \sqrt{S_y^2 \cdot (1 - R_{xy}^2)}. \tag{6.6}$$

in which S_y^2 is the criterion variance and R_{xy}^2 is the variance explained by the regression.

After calculating effect sizes within each sample, we separately accumulated the d-values for men and women. Two bare-bones meta-analyses were performed (i. e., correction of the observed variance for sampling error; Hunter & Schmidt, 2004). We chose the random-effects model as we expected effect size variations based on sample characteristics (Borenstein, Hedges, Higgins, & Rothstein, 2010). Significant Q-statistics in the fixed-effects tests of homogeneity should support this choice. The corresponding formula for the weighted average effect size was

$$\bar{d_j} = \sum w_{ji} d_{ji} / \sum w_{ji} \qquad (j = \text{men, women}; i = \text{sample}) \tag{6.7}$$

with w_i as the weight for the ith study. The inverse of the variance, which for one variable relationship is sample size divided by the variance of the target variable, was used as the weight (Lipsey & Wilson, 2001, p. 72). Since we had standardized effect sizes d (and not raw mean residuals) the standard deviation of the target variable was 1, so $w_i = n_i$. For a corresponding example, see Hunter and Schmidt (2004, p. 289). Further corrections for artifacts were not possible because there were no artifact information reported with regard to the mean residuals.

We calculated 95 % confidence intervals, 90 % credibility intervals, and tested the homogeneity of the effect sizes. The homogeneity test allows conclusions on whether the samples do share a common population effect size or not. We used the heterogeneity test (Q-test) according to Shadish and Haddock (2009) based on a fixed-effects model. Finally, we conducted moderator analyses to examine the source of heterogeneity.

Summarizing regression equations. Although the methodological literature on meta-analytic techniques is substantial, little attention has been paid to the issue of summarizing regression slopes and intercepts. This fact can be explained by several challenges (for a detailed discussion see Aguinis et al., 2010).

An overview of the existing methods for summarizing slopes was given by Becker and Wu (2007). They addressed the shortcomings of these methods by presenting a new multivariate generalized least squares approach. Anyhow, this method is also limited, because it requires knowledge of the covariances among slopes, which are rarely reported.

A new approach recommended using the semi-partial correlation (between predictor and criterion) as a partial effect size (Aloe & Becker, 2011; for an example, see Aloe & Becker, 2009). This method allows summarizing linear models with more than one predictor. In the present model, we have two predictors and an interaction term (see equation 6.1). To aggregate the interaction terms the corresponding t-statistic as well as the correlation between the interaction term and the test score is needed (given that, the test score and sex are related). Our identified studies rarely reported this information (especially the correlations). Moreover, some studies only reported the standardized regression weight for gender (e. g., Bridgeman & Wendler, 1991) and others reported the contribution to R^2 of intercept and slopes (e. g., Pennock-Román, 1994). This mixture of available information about regression equations is in line with Borneman's (2010) conclusion that "it is unlikely there are sufficient data in published manuscripts lying around for meta-analysis" (p. 225). Despite theoretically having the desired statistical properties, methods for aggregating regression equations could not be applied because the relevant studies did not report sufficient data. In order to still summarize the studies reporting regression equations, we created a descriptive summary.

6.3 Results

6.3.1 Gender-Specific Residuals

For each gender, we calculated separate mean effect sizes based on (a) studies that used admission test results as the sole predictor, and (b) studies that used a combination of admission test results and HGPA or UGPA as the predictor. We also calculated mean effect sizes across the studies, without the four large-scale studies ($N > 5,000$). The results were substantially the same.

Egger's regression test for funnel plot asymmetry (Sterne & Egger, 2005) is not significant for three of the four funnels (females: admission test as predictor $t = 1.23$, $p = .225$; females: admission test and HGPA/UGPA as predictor $t = 1.90$, $p = .063$; males: admission test and HGPA/UGPA as predictor $t = .93$, $p = .356$). Only the funnel of the effect sizes for males, based on the admission test as the sole predictor, reaches significance ($t = 2.52$, $p = .015$). Since the effect sizes for males and females are based on the exact same amount of unpublished and published studies, we think there is no substantial publication bias. Moreover, the asymmetric funnel is related to an unequal amount of men and women in some samples. There are two outliers in the funnel identified as asymmetric. Both samples show a very low percentage of men (less than 28 %). Following equation 6.3, residuals are not independent of sample sizes. Particularly a small proportion of one group within one sample can result in an extreme mean residual for this group. This is the case with these two outliers.

Admission test as predictor Table 6.1 shows the corresponding mean effect sizes for women ($d_{\text{female}} = .14$, indicating underprediction) and men ($d_{\text{male}} = -.16$, indicating overprediction). Because neither confidence nor credibility intervals overlap zero, the results can be generalized (Hunter & Schmidt, 2004). According to a fixed effect model the effect sizes are heterogeneous for women, $Q_{\text{fix}}(54) = 111.31$, $p < .001$ and homogenous for men $Q_{\text{fix}}(54) = 69.49$, $p = .076$.

Admission test combined with HGPA or UGPA as predictor For studies using admission tests and HGPA/UGPA as predictors, mean effect sizes are slightly smaller ($d_{\text{female}} = .11$ and $d_{\text{male}} = -.12$; see Table 6.1). It must be kept in mind that these results are not completely independent from the analyses of the studies that were using only admission tests as the predictor due to overlapping samples. Again, the confidence intervals as well as the credibility intervals do not include zero. The heterogeneity analyses reveal that the effect sizes are heterogeneous for women $Q_{\text{fix}}(50) = 76.75$, $p < .01$ and homogeneous for men $Q_{\text{fix}}(51) = 49.76$, $p = .523$.

Moderator analysis Effect sizes are heterogeneous for women based on admission tests as the sole predictor. To explain this effect size distribution for women we conducted analyses for potential moderators.

Test type is a significant moderator, $Q_{\text{between}}(5) = 76.97$, $p < .001$. The effect sizes for the different tests where more than one sample was available are displayed in Table 6.2. The underprediction of women's academic performance is negligible for the graduate admission tests, GRE and Medical College Admission Test (MCAT), and small to moderate for the undergraduate admission tests, SAT[19] ($d_{\text{female}} = .14$) and ACT ($d_{\text{female}} = .30$).

Separate prediction results for mathematics and verbal test sections were reported by four large-scale SAT studies. The mathematics section shows more underprediction of women ($d_{\text{female}} = .17$, $k = 4$, $N_{\text{female}} = 135,144$, 95 % CI [.15, .19], 90 % CRI [.15, .19]) than the verbal section ($d_{\text{female}} = .10$, $k = 4$, $N_{\text{female}} = 135,144$, 95 % CI [.09, .11], 90 % CRI [.10, .11]).

As shown in the previous analyses about the moderator test type, the study about the ACT (American College Testing Program, 1973) provides extreme over- and underprediction. At the same time, this study is by far the oldest one and the time span between predictor and criterion assessments is very short. We therefore tested the moderator variables, publication year and time span, with and without the ACT study. The results indicate the great influence of the ACT study. The significant influence of the moderators

[19]The analysis includes older SAT versions as well as the revised SAT version, which includes a writing component.

disappears if the ACT study is removed from the analyses. The influence of the remaining moderator variables predictor differences and criterion differences is not significant. Statistics for the moderator analyses are presented in detail in Table 6.3.

For the moderator variables, age and course choice, there was not enough data from the primary studies for the analysis.

6.3.2 Differences in Group Regression Equations

As described in the section *summarizing regression equations*, we present a descriptive summary of the studies offering group regression equations. Out of the included samples using admission tests as the sole predictor ($k = 20$, $N = 31,798$), 14 (70 %) show significant slope and/or intercept differences, which indicate differential prediction. Eight of the samples showing differential prediction underpredict women's performance and overpredict men's performance. One sample shows no clear direction of the effect. The other five samples neither report conclusions about over-/underprediction nor report the required statistics to derive the information.

Predictions using a combination of admission test and HGPA or UGPA ($k = 35$, $N = 51,436$) show differential prediction less often. In 16 of these samples (46 %), significant slope and/or intercept differences appear. Out of these samples, six underpredict women's performance, whereas one underpredicts men's performance. Unfortunately, the other nine samples do not report conclusions about overprediction and underprediction or the required statistics to derive the information.

Noticeably, the average sample size of studies reporting significant slope or intercept differences is higher ($N_{\text{mean}} = 2,032$) than the average sample size of the studies reporting no differences ($N_{\text{mean}} = 573$). This is not a surprise since significance depends, besides other factors, on sample size.

Table 6.1: Differential Prediction Effects for Women and Men

	Women					Men				
Predictor(s)	*k*	*N*	*d*	95 % CI	90 % CRI	*k*	*N*	*d*	95 % CI	90 % CRI
Admission test	55	154,162	.14	[.13, .16]	[.08, .21]	55	140,950	-.16	[-.17, -.15]	[-.20, -.13]
Admission test and HGPA/UGPA	51	220,321	.11	[.10, .12]	[.07, .14]	52	203,940	-.12	[-.12, -.11]	[-.17, -.06]

Note. Positive effect sizes indicate underprediction, whereas negative effect sizes indicate overprediction.

k = number of samples; CI = confidence interval; CRI = credibility interval.

Table 6.2: Differential Prediction Effects for Women moderated by Test Name

Test name	Studies	k	N_f	d_f	95 % CI	90 % CRI	Q_{within}	p
SAT	6	7	139,856	.14	[.13, .15]	[.12, .16]	5.67	.461
ACT	1	19	8,928	.30	[.25, .34]	[.23, .36]	21.85	.239
GRE	5	13	2,589	.03	[−.02, .08]	[−.11, .18]	4.85	.963
MCAT	1	14	1,312	.02	[−.02, .06]	[−.11, .15]	1.96	.999

Note. Studies = number of studies included; k = number of samples; whereas negative effect sizes indicate overprediction. k = number of samples; CI = confidence interval; CRI = credibility interval.

Table 6.3: Influence of Moderators on Differential Prediction Effects for Women

Moderator	k	β	R^2	p
Publication year	55	−.658	.43	<.001
Publication year [a]	36	.212	.04	.200
Predictor differences [b]	14	−.029	.00	.936
Criterion differences [b]	32	−.100	.01	.694
Time	44	−.344	.12	<.05
Time [a]	25	.314	.10	.085

Note. Studies that report insufficient data to code a particular moderator are omitted from that analysis; therefore, k fluctuates between analyses. Predictor and criterion differences are based on effect sizes, subtracting women's scores from the men's scores, respectively. Positive betas denote increases in women's effect size as the value of the predictor increases, whereas negative betas denote decreases in effect size as the value of the predictor increases. k = number of samples; time = time between admission test and criterion measure. R^2 = explained variance calculated conventionally following Lipsey and Wilson (2001).

[a] Analysis without the ACT study (American College Testing Program, 1973).

[b] Predictor differences were corrected for criterion differences and vice versa, if the required statistics were given. We also performed the analyses without the corrections; the results were essentially the same.

6.4 Discussion

The analysis of residuals shows that undergraduate and graduate admission tests underpredict women's academic performance ($d = .14$) and overpredict men's academic performance ($d = -.16$), on average. According to Cohen's (1988) classification, these effect sizes are less than small. This classification was an initial general attempt and not intended to be applied to every situation. Less than small underprediction may still have tangible consequences for admission decisions. Aguinis et al. (2005) showed that this occurs frequently in studies of differential prediction.

When the effect sizes are transferred onto a four-point grading scale (plugging in the mean standard deviation of residuals of the studies with the largest sample sizes), the academic performance of women is .11 points better than that predicted by the test. At the same time, men achieve grades that are .13 points worse than that predicted. In other words, with the same admission test result, women earn .24 points better grades than men do. The amount of underprediction and overprediction is smaller when admission tests are used in combination with HGPA/UGPA ($d_{\text{female}} = .11$, $d_{\text{male}} = -.12$)[20]. In fact, the academic performance of women is .08 points better and the academic performance of men is .09 points worse than predicted. Taken together, our research confirms the findings of Young and Kobrin (2001), who report a mean underprediction of women's performance of .06 grade points. However, our results also show that the differential prediction effect is almost twice as big if the admission test is used as the sole predictor.

Studies comparing regression equations yield similar results. Samples in which admission tests are used as the sole predictor show differential predic-

[20]This fact raises the question, whether HGPA or UGPA are biased in the opposite direction, that is, overpredicting women's academic performance. We analyzed the mean effect size of differential prediction for HGPA or UGPA for the included samples. The results show very small underprediction for women ($d_{\text{female}} = .07$, $k = 24$, $N_{\text{female}} = 144,383$, 95 % CI [.06, .09], 90 % CRI [.03, .12], $Q(23) = 50.99$, $p < .001$) and very small overprediction for men ($d_{\text{male}} = -.08$, $k = 24$, $N_{\text{male}} = 131,675$, 95 % CI [-.09, -.06], 90 % CRI [-.11, -.04], $Q(23) = 38.93$, $p < .05$). In short, HGPA or UGPA seems to be biased in the same direction as admission tests, but the magnitude is attenuated.

tion more often than those with a combination of admission test results and HGPA/UGPA (70 % versus 46 %). The prevalent direction of the effect is underprediction of women's academic performance. The number of studies that find no differential prediction is surprisingly small when compared to the number of studies that show group-specific residuals around zero. This might be because of publication bias, that is, the tendency for null results to remain unpublished. Further, almost all samples used undergraduate admission tests as a predictor, whereas the studies that show group-specific residuals around zero are mostly based on graduate admission tests.

6.4.1 Possible Reasons for the Underprediction of Women's Academic Performance

First, underprediction of women's performance remains an ongoing topic, as the differential prediction could not be reduced during the last decades though items are well reviewed for content fairness. The underprediction of women is further associated neither with test score differences (possibly indicating a bias in the test) nor with grade differences (possibly indicating some type of bias in the criterion; Meade & Fetzer, 2009). Consequently, disposing test score differences, by restructuring a test, does not necessarily reduce underprediction.

Different levels of over- and underprediction are rather related to different admission tests. Graduate admission tests indicate less of a problem with underprediction than undergraduate admission tests. This conclusion is consistent with the findings of Kuncel and Hezlett (2007). The underprediction linked to undergraduate tests might be explained by differences in *studying habits*. Women typically devote more effort to their academic work and show higher class attendance and greater academic motivation (Zwick, 2002). When accounting for such variables sex-specific differential prediction was reduced (Stricker, Rock, & Burton, 1993). As graduate students are a more *selective sample*, they possibly show more homogeneous personality characteristics and skills across the sexes than high school alumni.

Another explanation might be more course choice possibilities during un-

dergraduate studies. Unfortunately, there is not enough data to test the influence of course choice on differential prediction within the present study.

A further prominent explanation for the underprediction of women's academic performance is the influence of *stereotype threat* during the test execution. This means, women are under additional pressure that interferes with their test performance, because men are expected to outperform them on tests (e. g., Spencer, Steele, & Quinn, 1999; Steele, 1997). Still, the examination of stereotype threat in real world settings is difficult and is just at the beginning (Sackett, 2003; Sackett, Hardison, & Cullen, 2005). Given that we find differential prediction only in some admission tests makes it implausible that differential prediction is strongly associated with stereotype threat.

We also found differences between test components. The SAT verbal section shows less underprediction than the mathematics section. One explanation could be the *multiple-choice format* which is more prevalent for mathematical than for verbal sections. Men tend to perform better in multiple-choice formats than women, whereas women reach at least equal scores in free-response formats (Bridgeman & Lewis, 1994; Lindberg, Hyde, Petersen, & Linn, 2010). Lack of time might be responsible for these differences (Goldstein, Haldane, & Mitchell, 1990).

Summing up, our results indicate that the underprediction of females' academic performance is not related to test score differences or criterion differences. Moreover, especially undergraduate admission tests are prone to differential prediction. Sex differences of undergraduate students with regard to their study habits and motivational factors are promising explanations that call for future investigations.

6.4.2 Strengths and Weaknesses of Methods Measuring Differential Prediction

Testing for differences in regression lines Analysis of differences in regression lines is easy to illustrate and has been used for years. However, most of the relevant studies failed to report the information required to aggregate the results with meta-analytic techniques. Another pitfall concerns

the intersection of regression lines. If regression lines intersect at a low predictor level, the intercept test can reveal underprediction of women, while the test score range containing most applicants overpredicts women (Schmidt & Hunter, 1982). To avoid this problem it is recommended to center the predictor variable or to define the areas where the group-specific regression lines differ. Only few studies implemented these recommendations. Therefore, it is possible that the results are artifacts and do not allow conclusions about the actual research questions.

Finally, finding significant differences between intercepts or slopes depends on sample size. In contrast to the meta-analytic approach, a descriptive overview cannot take this problem into account.

Reporting group-specific residuals Reporting residuals helps to communicate test properties to a lay audience. Unstandardized mean residuals can be easily interpreted as the average deviation from the common regression line in the unit of the criterion scale. The residual method can be applied to large-scale studies and residuals can be transformed into effect sizes, which can be aggregated in meta-analysis. Despite these advantages, the method has its limitations.

The mean residual for women is not independent from the male residual and their sample sizes, respectively (see equation 6.3). As a consequence, minorities reach more extreme mean residuals than the corresponding majority. When a common regression line is inappropriate, the analysis of residuals can be misleading as well. This is the case when slopes of group-specific regression lines have different algebraic signs, so that the lines intersect near the mean of the predictor (Norborg, 1984). In other words, one-half of each group (i. e., the upper- or lower-half on the test score scale) is overpredicted, whereas the other half is underpredicted. The coexistence of overprediction and underprediction remains undetected because both group mean residuals are zero.

Methodological Conclusions With regard to future meta-analysis of differential prediction results, we recommend that all relevant information in

primary studies should be reported. Additionally, the scatter plots of the group-specific regression lines should be inspected to determine the curve progressions, especially potential intersections. In some cases, it might be helpful to further provide residual results for different predictor regions, for example, around the *cut-off point* used for admission. Additionally, the variance-covariance matrix should be provided to enable the meta-analysis of regression slopes (Borneman, 2010) as well as correlations (Aloe & Becker, 2011).

6.4.3 Final Conclusion

The present meta-analysis shows that admission tests underpredict women's academic performance and overpredict men's academic performance to a small but consistent extent. This conclusion holds true for older as well as for newer tests and is not related to predictor or criterion differences. Particularly undergraduate admission tests are more prone to over- and underprediction effects than graduate tests. Future research should build on these results. We suggest to focus on sex differences in non-cognitive factors like study habits and motivational factors of undergraduate students rather than on test or criterion differences.

7 General Discussion

The extent of sex-specific differential prediction was analyzed in Studies 1, 3 and 4, while Study 2 focused on the interplay of construct and criterion validity of scholastic aptitude tests. Each study provided an answer to a hitherto underresearched question.

Study 1 showed that the sex-specific differential occurs in German student populations. Numeric subtest scores tend to favor men whereas verbal subtest scores tend to reduce this underprediction—or even convert it into an overprediction of college grades. This result is illustrated in Figure 7.1.

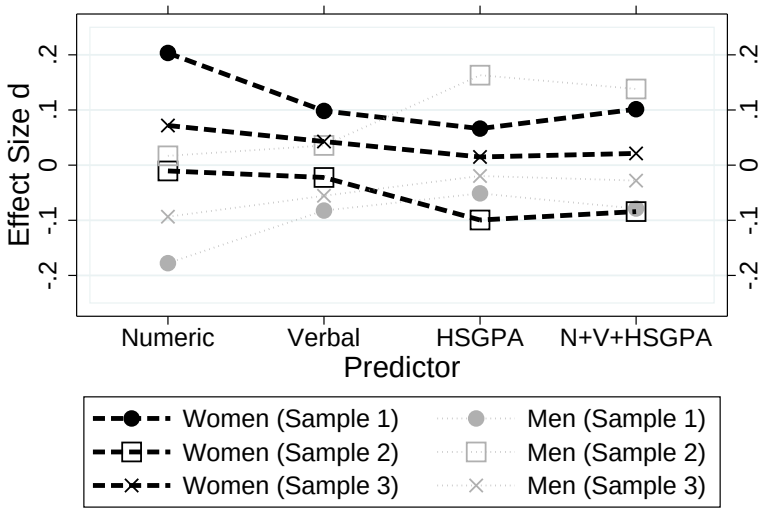


Figure 7.1: Effect size d of residuals by predictor type for each sample in Study 1. Negative residuals indicate overprediction; positive residuals indicate underprediction; V = verbal, N = numeric, HSGPA = high school grade point average (see Study 1 in Section 3 for sample details).

Study 2 highlighted the role of intelligence facets in the prediction of academic success and provided cognitive insights into the construct validity of two German scholastic aptitude tests. Science courses appear to be not all about numeric reasoning, but—particularly for female freshman students—also about verbal reasoning abilities.

Study 3 extended the scope of sex-specific differential prediction, using success at work two years after graduation as outcome. Here, personal interests, occupational status, math grades, and conscientiousness are the best predictors. Sex-specific income differences remain even after controlling for socio-economic and motivational factors.

Study 4 showed how differential prediction findings can be aggregated. The underprediction of women's academic performance is mainly an issue for undergraduates. Tests for graduate courses, on the other hand, do not show differential prediction.

7.1 Differential Prediction

College admission tests are a selection tool that runs both ways: academic institutions attempt to select the most promising applicants whereas prospective students try to find the college that meets their expectations and needs best. Biased test results will not just affect the selection procedure of colleges. They are also likely to keep some students from applying in the first place who would have easily succeeded in college. Often, the majority of this group of people is female (cf. Young & Kobrin, 2001). The meta-analysis presented in Study 4 supports this notion.

The underprediction of women's academic performance has the undesired effect that potentially successful female applicants who score close below the cut-off point are denied college admission. Population statistics indicate that women are not barred from going to university (Snyder & Dillow, 2012), but they may end up at less prestigious universities and choose different majors (Ceci et al., 2009).

Another issue may be that potentially successful female applicants who score close below the cut-off point (erroneously) take their scores as a sign that they are unlikely to succeed in college. This could keep them from attending college on their own volition and choosing an educational path outside of university (Atkinson & Geiser, 2009). Support for this effect is weak (Holden, 1989) and it could well be the other way round: men with moderate aptitude are more likely to be admitted to college—and fail. The U. S.

American data are in line with this scenario. Drop-out rates for 2010 were estimated to be 8.5 % for men and 6.3 % for women (Snyder & Dillow, 2012). So the effect of differential prediction is—at least partially—neutralized by differential validity. Fewer potentially successful women are initially admitted, but lower validity coefficients for men suggest that more men fail to succeed.

One of the short-comings of the meta-analysis in Study 4 is that it includes mostly U. S. American studies. Can the results be translated to Germany?

The differential prediction analysis of the data used in Study 2 reveals that scholastic aptitude test scores tend to under-predict women's FYGPA, whereas high school grades show an opposite effect; only the combination of both predictors in a multiple regression model yields fair predictions (see Figure 7.2 for a summary). This result resembles data from the German admission test for medical colleges (TMS; Nauels & Meyer, 1997). Using HSPGA in addition to test scores does not only diminish the differential prediction of the admission test, it counterbalances the differential effect almost perfectly.

Further analysis suggests that differential prediction associated with tests of scholastic aptitude is particularly pronounced in competitive settings where only few (i. e., the very best) applicants are selected (cf. F. Fischer, Schult, & Hell, in revision). This finding is at odds with recent U. S. American data, which shows very small slope differences (Mattern & Patterson, 2013).

The differential prediction results from Study 1 support the notion that HSGPA is generally less favorable for men, although d_{male} is not always positive (see Figure 7.1). The differential prediction associated with intelligence test scores is heterogeneous, but on average in line with the findings from college admission tests[21].

Two studies are certainly no match for a meta-analysis that covers many tests, samples, and decades. My current conclusion is that in Germany,

[21]In Study 2, the correlations between reasoning scores and scholastic aptitude test scores are moderate (economics: $r_{\text{verbal,SFBT}} = .34, r_{\text{numeric,SFBT}} = .54$; science: $r_{\text{verbal,SFBT}} = .46, r_{\text{numeric,SFBT}} = .50$), suggesting a sizeable overlap between measures of mental ability and scholastic aptitude that justifies the use of intelligence tests in Study 1.

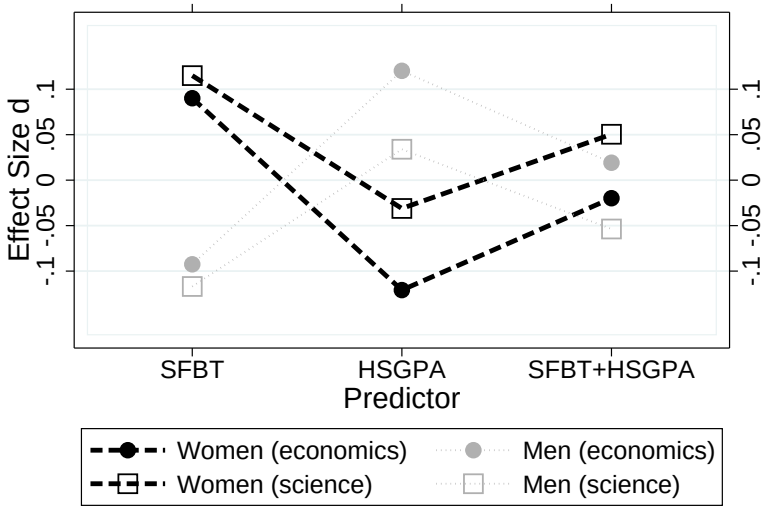


Figure 7.2: Effect size d for residuals by predictor type for each field of study in Study 2. Negative residuals indicate overprediction; positive residuals indicate underprediction; SFBT = scholastic aptitude test score, HSGPA = high school grade point average (see Study 2 in Section 4 for sample details).

the differential prediction associated with scholastic aptitude tests is less pronounced than in the meta-analysis. The combination of HSGPA and aptitude test scores appears to yield the most favorable predictions in terms of both gender fairness and overall validity[22]. While it may be tempting to recommend admission tests unconditionally based on these findings, there are some issues that need to be addressed—most pressingly the socio-economic fairness of scholastic aptitude tests. Other reservations concern the selectivity of validity studies and to the lack of clarity regarding what GPAs measure and whether that measurement differs between teachers, majors, and countries.

[22]In Study 2, the incremental validities of aptitude test scores are $\Delta r = .05$ (economics) and $\Delta r = .03$ (science), respectively. Overall validities of HSGPA and test scores combined are $r = .63$ (economics) and $r = .47$ (science), respectively.

7.2 Explanations for Sex-Related Predictive Bias

So far, this discussion has focused on the empirical college admission data and practical actions that can be deduced from them. Surprisingly little is known about the underlying causes of biased predictions. Some leads appear to be more promising than others, but frankly, it is too early to arrive at any definite conclusions.

Mean differences in either test scores or college grades are prominent suspects. Tests could contain material that men are more familiar with; mean differences in grades could indicate a biased assessment by the teacher rather than inherent group differences. A proper test of these ideas requires either an infallible predictor or an infallible outcome—something which does not exist in college admission testing (Linn, 1984). Consequently, mean differences alone do not say anything about the fairness of a testing procedure (Linn, 1990a; Meade & Tonidandel, 2010). A straightforward way to identify items that may discriminate against a particular subgroup is the analysis of DIF, which can help identify items that, for example, men are more familiar with than women (Curley & Schmitt, 1993; Osterlind & Everson, 2009). There is no systematic effect of mean test score differences according to the moderator analysis of Study 4. Mean differences in CGPA are also not related to differential prediction. There is some evidence suggesting that courses with lenient grading styles are more frequently attended by women (Alon & Gelbgiser, 2011; Berry & Sackett, 2009). Still, the majority of studies on differential prediction use proximal criteria which are assessed while there is little room for differential course choices. Sex differences in self-imposed workload are equally unlikely[23]. There is, however, a distinct difference between undergraduate and graduate tests in the meta-analysis.

The decision to become a graduate student depends on cognitive ability as well as on motivation—previous educational attainment is usually a prerequisite, but without interest in a specific field of study other careers might be more appealing. This self-selection process is usually found in validity

[23]During their first year, the students in Study 2 took exams for 50 ECTS points on average. The workload of men and women did not differ ($t(640) = 0.25; p = .80$).

studies of graduate admission tests (Kuncel et al., 2001).

7.2.1 Sex Differences in Interests

Sex differences in cognitive ability may come into play at highly selective elite institutions, but appear to have little impact elsewhere (Ellis et al., 2008). Meanwhile, some of the largest psychological sex-differences can be found in vocational interests (Su et al., 2009). Similarly, men and women tend to choose different majors (see Table 5.2 and also Bridgeman et al., 2000). Validities differ between fields of study: predictive validity coefficients are higher in science and math majors, where there is a male majority. Still, the sex-specific differential validity in favor of women persists within each group (Bridgeman et al., 2000).

Evaluating science courses separately leads to a reduction of sex-specific differential prediction (Bridgeman et al., 2000; Elliott & Strenta, 1988). The interplay of self-selection processes and curricular differences is a promising field for future research.

More detailed findings are already available from studies of success at work (e. g., S. Anger & Heineck, 2010; Hinz & Gartner, 2012): About half of the initial gender pay gap observed in income data can be explained by sex differences in interests and job status. Unlike income, job satisfaction is unrelated to a person's interest and sex (see Study 3). It is conceivable that women choose careers that are associated with lower wages and fewer promotion opportunities in favor of jobs—and educational tracks—they find more satisfying (cf. A. E. Clark, 1997; Fietze et al., 2010; Kaiser, 2005).

Finally, the personal attributes associated with highly succcessful students appear to be distributed unevenly between men and women: on average, men are willing to devote more working hours to their careers, which makes them more likely to excel in their chosen field (Lubinski & Benbow, 2007). Apparently, men tend to be more willing to pursue a single goal whereas women tend to embrace a broader set of challenges. The next section shows that this is also the case when it comes to individual tasks.

7.2.2 Sex Differences in Dealing with Complexity

Another promising lead is task complexity. When it comes to complex problem solving, men are more likely to apply the strategy to explore one thing at a time whereas women prefer a more holistic approach (Wüstenberg, Greiff, Molnar, & Funke, submitted). There is a similar pattern for complex quantitative tasks, where female students are more likely to use strategies that use additional cues (Spelke, 2005). This approach may be a disadvantage for women in well-defined test settings; but when faced with the various challenges of being a student at university, a holistic approach might be more efficient and, eventually, more successful. Furthermore, the knowledge and skills acquired at university defy simplicity and call for a broad set of cognitive abilities, personality traits, and process-related aptitude, instead (Stemler, 2012). Already in school, young women tend to have superior social skills and show less disruptive classroom behavior, which may eventually facilitate their academic careers (Buchmann, DiPrete, & McDaniel, 2008). So possibly, the very thing that keeps women from getting the same average scores in college admission tests, is what helps them outperform men at university. The sex-specific construct validity of scholastic aptitude tests reported in Study 2 is in line with this hypothesis. A different set of intelligence facets is associated with women's aptitude test performance.

Conventional admission tests often contain many shorter items rather than a few elaborate questions, because time is limited (due to economic and logistic constraints) and more items tend to make a test more reliable. These items usually have a pre-defined correct answer, which can be found most easily by convergent thinking (Dollinger, 2011).

Essay writing parts were introduced to the SAT and the ACT in 2005. The SAT Writing subtest (SAT-W) contains an essay that is scored holistically; readers are asked to judge the total impression of an essay rather than distinct factors (Camara, 2003; Kobrin, Deng, & Shaw, 2011). The predictive validity of the SAT-W is similar to that of the SAT Critical Reading (SAT-CR) subtest and the SAT Mathematics (SAT-M) subtest, but it provides very little incremental validity ($\Delta r \leq .02$) over SAT-CR and SAT-M (Norris,

Oppler, Kuang, Day, & Adams, 2006). On average, women get higher scores than men on the SAT-W, so it is no surprise that the underprediction of women's FYGPA is smaller for the SAT-W than for the SAT-CR and the SAT-M (Mattern et al., 2008; B. F. Patterson & Mattern, 2011).

It seems as if men are one-trick ponies who are better off in reasoning tests that require convergent thinking. Yet, women appear to be better equipped to handle the ambiguous reality of studying at university with its manifold challenges. Motivational aspects that are relevant for scholastic success tend to be more prominent among young women (e. g., self-discipline; Duckworth & Seligman, 2006). Women also outperform men in high school, again a setting that is a complex collection of various subjects and assessments that require a heterogeneous set of skills and often go beyond short, clear-cut tasks (Buchmann et al., 2008). This notion persists beyond graduation: the link between high mathematics competence and high verbal competence is stronger in women than in men. Thus, women have more flexibility in the job market and can choose from a wider array of professions (Ceci et al., 2009).

If the aim is to reduce sex-specific differential prediction, future tests of scholastic aptitute might benefit from tapping into the mastery of real-life challenges and divergent tasks opposed to convergent problem solving that requires only few different skills. In this sense, introducing essay tasks like the SAT-W is a step into the right direction.

7.3 Where Do We Go From Here?

7.3.1 The Psychometrics of Grading

Translating the idea of test fairness to the prediction of income provides not only insights regarding the gender pay gap; it also uncovers a weak spot in college admission testing: the psychometrics of grading.

With a criterion as poorly understood as GPA[24], causes for differential

[24]Income—as indicator of job success—is also difficult to deal with given its manifold dependencies (Fietze et al., 2010; Judge et al., 1999). Consequently, the issue discussed here concerns the differential prediction of wages, as well.

prediction can easily hide in measures of academic performance. College teachers do not necessarily have any diagnostic training. This opens the door for biased grades (Birkel & Birkel, 2002; Elliott & Strenta, 1988; Waugh & Gronlund, 2013). Aggregating individual grades to GPAs helps harmonizing the outcome measures (Berry & Sackett, 2009). Still, subject-specific grading styles in connection with sex-specific subject choices might lead to differential prediction (Ceci et al., 2009).

The SOEP data presented in Study 3 show that the effects of sex-specific choices of majors persist at work throughout the first years after graduation. This is in line with the large differences between the vocational interests of men and women discussed earlier (Su et al., 2009).

7.3.2 Opportunities for Future Studies

The next best opportunity to study the relationship between choice of major and academic achievement will be the German National Education Panel Study (NEPS; Blossfeld, Roßbach, & von Maurice, 2011). Data collection started in 2010 with—among others—a cohort of 9th graders who are followed on through high school and, eventually, university. Another cohort comprises freshman students at university. Extensive measurements of competencies are assessed along the way, so that differential effects of educational choices and sex differences can be studied independently of grade points.

The testing industry is surprisingly eager to publish research reports regarding fairness issues. Most of their large-scale data are produced during the actual admission process. Validity studies then assess additional criterion variables in order to provide insights regarding differential validity and differential prediction for various subgroups (e. g., ACT, 2012; B. F. Patterson et al., 2009; Shaw et al., 2012). These data sets can be used to study a variety of relevant topics, for example the influence of grading styles on test fairness (Berry & Sackett, 2009) and the sex-specific interplay of abilities (Ceci et al., 2009). Unfortunately, the raw data are scattered across institutions and are not readily available as scientific use files. This may impede the formation of an integrative outlook on factors that affect a test's validity and fairness.

7.3.3 The Costs of College Admission Testing

The amount of differential prediction that is commonly reported for undergraduate admission procedures (see Study 4) may be not worrying, but it certainly looks odd. Why should an institution use a selection tool that yields biased predictions? From a psychometric point of view, the answer is incremental validity. Combining HSGPA and admission test scores does increase the amount of differential prediction, that is, the relative amount of successful graduates among the chosen applicants; but at the same time the total number of correctly identified successful students increases due to the predictive power of the test scores. Thus, the total number of successful women who are admitted increases. Of course, the number of successful men who are admitted increases even more. The bottom line is a paradox (see Holden, 1989): in order to be able to select more successful female students, one might have to put up with the differential prediction bias associated with tests of scholastic aptitude.

Unfortunately, it is not possible to explore all possible scenarios, because there are too many variables at play: test scores, high school grades, college grades, their intercorrelations, and, finally, possible sex differences associated with each of these variables and parameters. Research funding is limited. Looking at how grades work is a better way to approach to differential prediction than studying arbitrary moderator variables, which—at worst—lack a clear conceptual background. As shown in Studies 1 and 2, content domains of reasoning ability are also promising candidates for future research, especially because abilities—unlike many noncognitive variables—cannot be faked in a testing situation.

There is another money issue: developing and administering tests is expensive. It certainly helps that institutions like the College Board and ACT are not-for-profit organizations (for a discussion of cost distribution see also Trost, 2005). Still, the increase in correctly admitted women may literally come at a high price. Test scores improve the validity of the admission decision, but the costs of taking the test can have an adverse effect on the willingness of people with a low socio-economic background to ap-

ply for a highly selective institution or to pursue a college education in the first place. In Germany, the correlation between children's socio-economic background and their reading performance is already larger than .4 (OECD, 2010). There are some educational differences between federal states, but HSGPA (Abiturnote) is already a rather powerful predictor. College admission tests are more deeply engrained in the educational landscape in the U. S.[25], where high school grades are subject to the policies of local school districts and admission tests are a substitute for a national curriculum (Linn, 1990b; Mattern, Shaw, & Kobrin, 2011). Social disparities in education are only slightly better than in Germany (OECD, 2010). Admission departments have to choose the lesser of two evils. This decision must be informed by administering statistical as well as social and political sensitivity.

The costs and benefits of college admission testing are summarized in Table 7.1 along with test fairness implications.

Table 7.1: The costs and benefits of college admission testing (cf. Trost, 2005)

Level	Costs	Benefits	Fairness implications
Individual	Fees, effort, time	Feedback, improved selection quality	Adverse impact due to underprediction
Institution	Development and application of tests	More successful students, standardized scores	Restricting diversity
Society	Regulation of admission testing	Tapping the population's full potential	Ethical and legal issues

[25]The incremental validity of SAT scores over HSGPA is actually the largest for students whose parents have a low level of education (Shaw et al., 2012), yet another sign that test fairness is a complex issue.

7.4 Conclusion

It is important to not rely solely on the numbers presented in validity studies. Instead, a proper admission policy that accounts for the various aspects of test fairness discussed in this thesis is required. As Darlington (1971, p. 71) wrote more than 40 years ago:

> If a conflict arises between the two goals of maximizing a test's validity and minimizing the test's discrimination against certain cultural groups, then a subjective, policy-level decision must be made concerning the relative importance of the two goals.

Transparent validity studies should be performed and published on a regular basis. A sufficiently large data base of test scores and associated test taker variables (especially academic performance and characteristics that are susceptible to discrimination) will help to tell a momentary bias increase from a systematic deterioration of fairness in college admission testing.

Test fairness is a moving target that varies between target populations, after all. A particular instrument might work with one field of study, but not with another. A subtest might provide a fair prediction at one point in time, but not five years later. The present thesis provides practical solutions to biased admission tests.

When swift action is needed, weighting subtest scores is recommended in order to counterbalance differential prediction effects—usually by limiting the impact of numerical subtests when the underprediction of women's academic performance needs to be addressed (cf. Figure 7.1). This adjustment is still possible after the testing sessions are over.

When it comes to developing admission tests, including items that cover a broad set of facets of cognitive ability keeps options open. The general factor of cognitive ability, g, has itself a high predictive validity, but it does not cover the entire range of abilities which is usually comprised by scholastic aptitude (see Study 2 and also Coyle & Pillow, 2008). Complexity appears to be a promising candidate that can help explain sex-specific differential validity and differential prediction.

Test fairness must be taken into account along with criterion validity. And since test fairness comes in various forms, college admission procedures must be willing to adapt to changes in society, politics, and culture.

8 Abstract

The present thesis answers several open questions regarding the gender fairness of scholastic aptitude tests and provides practical advice how to assess test fairness and minimize predictive bias.

There are several reasons to use aptitude tests in the college admission process: they offer standardized scores, provide incremental validity over high school records, and may influence the educational decisions of applicants. Despite the usefulness of these tests in practice, their construct validity, the reasons for group differences and other psychometric aspects usually remain unclear. Consequently, a closer look at the fairness of admission tools reveals many gray areas where improper test use and imprecise conceptualization cannot be easily distinguished.

Test fairness in a narrow, psychometric sense is based on the lack of systematic bias. Three types of bias are generally distinguished: differential item functioning (DIF; an item is more difficult (or easier) for a particular subgroup after controlling for the ability it is supposed to measure), differential validity (different criterion validities for subgroups), and differential prediction (performances of subgroups are systemically underpredicted (or overpredicted)).

Four studies have been conducted to shed light on the extent and possible explanations of sex-specific bias associated with scholastic aptitude testing and the prediction of academic and vocational performance. In the first two studies, special attention was given to the role of intelligence facets, because general mental ability (g) and scholastic aptitude overlap conceptually—reasoning is among the constructs assessed by most college admission tests—and are highly correlated.

Study 1 provides a detailed look at the situation in Germany. Three student samples show various levels of differential prediction. Across all samples, mathematical reasoning yields the most favorable predictions for men (i. e., men's college grades are overpredicted). High School Grade Point Average (HSGPA; "Abiturnote"), on the other hand, is the least favorable predictor for men's academic performance, although it still underpredicts

women's performance in two of the samples.

Study 2 explores the construct validity of two German tests of subject-specific scholastic aptitude. The link between intelligence and aptitude test score is confirmed. Small sex differences in validities suggest a stronger relationship between verbal reasoning and scholastic aptitude for women than for men.

Study 3 broadens the scope by looking at the careers of university students two years after their graduation. Valid predictors for success at work include personal interests, occupational status, math grades, and conscientiousness. The gender pay gap remains even after controlling for socio-economic status and motivational factors.

Study 4 demonstrates the aggregation of differential prediction findings with meta-analytical methods. The underprediction of women's college grades by aptitude tests can be reduced (but not eliminated) by using HSGPA and test scores as predictors. Graduate tests do not show differential prediction.

Based on these findings, two promising explanations for differential prediction are scrutinized. On the one hand, sex differences in vocational interests exist which are associated with choice of major and career paths. On the other hand, women appear to approach academic challenges in a more holistic way than men, which interferes with their admission test performance, but facilitates their academic performance, eventually.

Although some topics still need further attention (e. g., construct validity of grades, availability of large-scale data sets, socio-economic consequences of admission testing), my findings clarify the psychometric properties of scholastic aptitude tests and provide immediate suggestions for weighting subscales in order to maximize gender fairness.

9 Zusammenfassung

Diese Arbeit beantwortet einige offene Fragen zur Genderfairness von Studierfähigkeitstests und bietet praktische Hinweise bezüglich der Erfassung von Testfairness und der Minimierung von Vorhersageverzerrungen.

Es gibt mehrere Gründe, beim Hochschulzulassungsverfahren Fähigkeitstests einzusetzen: Sie liefern standardisierte Ergebnisse, haben inkrementelle Validität über Schulnoten hinaus und können den Bewerberinnen und Bewerbern Feedback liefern, das ihnen bei der Wahl ihres Bildungswegs hilft.

Obwohl die Tests in der Praxis hilfreich sein können, sind ihre Konstruktvalidität, die Hintergründe von Gruppenunterschieden und weitere psychometrischen Aspekte meist immer noch unklar. Entsprechend entdeckt man beim genaueren Betrachten der Fairness von Auswahlinstrumenten zahlreiche Grauzonen, bei denen sich unangebrachte Testanwendung und ungenaue Konzeptualisierung nur schwer voneinander unterscheiden lassen.

Testfairness im engeren, psychometrischen Sinn bezieht sich auf die Abwesenheit von systematischen Verzerrungen. Gewöhnlich werden drei Arten von Verzerrungen unterschieden: „differential item functioning“ (DIF; eine Aufgabe ist für eine Teilgruppe leichter (oder schwerer) unabhängig von der zu messenden Fähigkeit), differenzielle Validität (gruppenspezifische Kriteriumsvaliditäten) und differenzielle Prognose (Unter- bzw. Überschätzung der Leistung einer Subgruppe).

Vier Studien wurden durchgeführt, um die geschlechtsspezifischen Verzerrungen im Zusammenhang mit Studierfähigkeitstests und der Vorhersage von Studien- und Berufserfolg näher zu beleuchten. Bei den ersten beiden Studien liegt das Hauptaugenmerk auf Intelligenzfacetten, da der Generalfaktor der Intelligenz (g) und Studierfähigkeit sich konzeptuell überschneiden – logisches Schlussfolgern gehört zu den Konstrukten, die in den meisten Studierfähigkeitstests erfasst werden – und hoch miteinander korrelieren.

In Studie 1 wird ein genauerer Blick auf die Situation in Deutschland geworfen. In drei studentischen Stichproben fällt die differenzielle Prognose jeweils etwas anders aus. Über alle Stichproben hinweg liefert mathematisches Schlussfolgern die günstigste Vorhersage für Männer, d. h. eine

Überschätzung der Studienleistung. Die Abiturnote dagegen wirkt sich am ungünstigsten auf die Vorhersage für Männer aus, obwohl die Leistung von Frauen in zwei der Stichproben auch hier immer noch unterschätzt wird.

In Studie 2 wird die Konstruktvalidität von zwei fachspezifischen deutschen Studierfähigkeitstests untersucht. Der vermutete enge Zusammenhang von Intelligenz und Studierfähigkeitstestergebnis wird dabei bestätigt. Kleine Geschlechtsunterschiede bei den Validitäten deuten auf eine stärkere Verbindung zwischen verbalem Schlussfolgern und Studierfähigkeit bei Frauen hin.

In Studie 3 werden die Berufswege von Universitätsabsolventinnen und -absolventen zwei Jahre nach ihrem Abschluss betrachtet. Zu den validen Berufserfolgsprädiktoren gehören persönliches Interesse, Berufsstatus, Mathematiknoten und Gewissenhaftigkeit. Die geschlechtsspezifische Lohnlücke kann durch die statistische Korrektur für sozio-ökonomischen Status und motivationale Faktoren nicht geschlossen werden.

In Studie 4 wird die Aggregation von Befunden zur differenziellen Prognose mittels meta-analytischer Methoden demonstriert. Die Unterschätzung der Studiennoten von Frauen durch Fähigkeitstests kann durch die Prädiktorenkombination von Schulnoten und Testwerten reduziert, aber nicht beseitigt werden. Bei Tests für fortgeschrittene Studiengänge zeigt sich dagegen keine differenzielle Prognose.

Ausgehend von diesen Befunden werden zwei aussichtsreiche Erklärungen für die differenzielle Prognose eingehend untersucht. Einerseits hängen Geschlechtsunterschiede beim fachlichen Interesse mit der Studienfachwahl und dem anschließenden Karriereweg zusammen. Andererseits scheinen Frauen akademische Herausforderungen holistischer anzugehen als Männer. Das kann zwar die Zulassungstestleistung beeinträchtigen, erleichtert aber die Bewältigung des anschließenden Studiums.

Obwohl einige Themen noch mehr Aufmerksamkeit benötigen (z. B. die Konstruktvalidität von Noten, die Verfügbarkeit von großen Datensätzen, die sozio-ökonomische Konsequenzen von Zulassungstests), verdeutlichen meine Ergebnisse die psychometrischen Eigenschaften von Studierfähigkeitstests und liefern unmittelbare Empfehlung für die Gewichtung von Subskalen zur Maximierung der Genderfairness.

References

References marked with an asterisk indicate studies included in the meta-analysis.

Abedalaziz, N. (2010). A gender-related differential item functioning of mathematics test items. *International Journal of Educational and Psychological Assessment*, *5*, 101–116.

Abele, A. E., & Spurk, D. (2009). The longitudinal impact of self-efficacy and career goals on objective and subjective career success. *Journal of Vocational Behavior*, *74*, 53–62. doi: 10.1016/j.jvb.2008.10.005

Abele-Brehm, A. E., & Stief, M. (2004). Die Prognose des Berufserfolgs von Hochschulabsolventinnen und -absolventen. *Zeitschrift für Arbeits- und Organisationspsychologie*, *48*, 4–16. doi: 10.1026/0932-4089.48.1.4

ACT. (2012). *2011–2012 fairness report for the ACT tests.* Iowa City, IA: ACT. Retrieved from `http://www.act.org/research/researchers/pdf/AAP-FairnessReport.pdf`

ACT. (2013). *Newsroom: Other frequently asked questions about the ACT.* Retrieved from `http://www.act.org/newsroom/factsheets/view.php?p=160`

Aguinis, H. (2004). *Regression analysis for categorical moderators.* New York, NY: Guilford Press.

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, *90*, 94–107. doi: 10.1037/0021-9010.90.1.94

Aguinis, H., Boik, R. J., & Pierce, C. A. (2001). A generalized solution for approximating the power to detect effects of categorical moderator variables using multiple regression. *Organizational Research Methods*, *4*, 291–323. doi: 10.1177/109442810144001

Aguinis, H., Culpepper, S. A., & Pierce, C. A. (2010). Revival of test bias research in preemployment testing. *Journal of Applied Psychology*, *95*, 648–680. doi: 10.1037/a0018714

Aguinis, H., & Smith, M. A. (2007). Understanding the impact of test validity and bias on selection errors and adverse impact in human resource selection. *Personnel Psychology*, *60*, 165–199. doi: 10.1111/j.1744-6570.2007.00069.x

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.

Allen, J., Robbins, S., Casillas, A., & Oh, I.-S. (2008). Third-year college retention and transfer: Effects of academic performance, motivation, and social connectedness. *Research in Higher Education*, *49*, 647–664. doi: 10.1007/s11162-008-9098-3

Aloe, A. M., & Becker, B. J. (2009). Teacher verbal ability and school outcomes: Where is the evidence? *Educational Researcher Behavioral Statistics*, *38*, 612–624. doi: 10.3102/0013189X09353939

Aloe, A. M., & Becker, B. J. (2011). Advances in combining regression results in meta-analysis. In M. Williams & W. P. Vogt (Eds.), *The SAGE handbook of innovation in social research methods* (pp. 331–352). London, England: Sage.

Alon, S., & Gelbgiser, D. (2011). The female advantage in college academic achievements and horizontal sex segregation. *Social Science Research*, *40*, 107–119. doi: 10.1016/j.ssresearch.2010.06.007

*American College Testing Program. (1973). *Assessing students on the way to college: Vol. 1. Technical report for the ACT assessment program*. Iowa City, IA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (4th ed.). Washington, DC: American Educational Research Association.

Anger, C., & Schmidt, J. (2010). Gender Pay Gap: Gesamtwirtschaftliche Evidenz und regionale Unterschiede. *IW-Trends*, *37*, 1–15.

Anger, S., & Heineck, G. (2010). Cognitive abilities and earnings – first evidence for Germany. *Applied Economics Letters*, *17*, 699–702. doi: 10.1080/13504850802297855

Atkinson, R. C., & Geiser, S. (2009). Reflections on a century of college

admissions tests. *Educational Researcher*, *38*, 665–676. doi: 10.3102/0013189X09351981

Bai, C., & Chi, W. (2011). *Determinants of undergraduate GPAs in China: College entrance examination scores, high school achievement, and admission route.* MPRA Paper No. 32797. Retrieved from `http://mpra.ub.uni-muenchen.de/32797/`

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment*, *9*, 9–30. doi: 10.1111/1468-2389.00160

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannan, R. (1978). Testing for fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology*, *31*, 233–241. doi: 10.1111/j.1744-6570.1978.tb00442.x

Becker, B. J. (2000). Multivariate meta-analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499–525). San Diego, CA: Academic Press. doi: 10.1016/B978-012691360-6/50018-5

Becker, B. J., & Wu, M.-J. (2007). The synthesis of regression slopes in meta-analysis. *Statistical Science*, *22*, 414–429. doi: 10.1214/07-STS243

Bejar, I. I., & Blew, E. O. (1981). Grade inflation and the validity of the Scholastic Aptitude Test. *American Educational Research Journal*, *18*, 143–156. doi: 10.3102/00028312018002143

Bergmann, C., & Eder, F. (2005). *Allgemeiner Interessen-Struktur-Test mit Umwelt-Struktur-Test (AIST-R/UST-R) – Revision.* Göttingen, Germany: Beltz Test.

Berry, C. M., & Sackett, P. R. (2009). Individual differences in course choice result in underestimation of the validity of college admissions systems. *Psychological Science*, *20*, 822–830. doi: 10.1111/j.1467-9280.2009.02368.x

Birkel, P., & Birkel, C. (2002). Wie einig sind sich Lehrer bei der Aufsatzbeurteilung? Eine Replikationsstudie zur Untersuchung von Rudolf Weiss. *Psychologie in Erziehung und Unterricht*, *49*, 219–224.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). *Education as a lifelong process: The German National Educational Panel Study (NEPS).* Wiesbaden, Germany: Springer VS.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*, 97–111. doi: 10.1002/jrsm.12

Borneman, M. J. (2010). Using meta-analysis to increase power in differential prediction analyses. *Industrial and Organizational Psychology*, *3*, 224–227. doi: 10.1111/j.1754-9434.2010.01228.x

Bortz, J., & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4th ed.). Heidelberg, Germany: Springer.

Bowman, N. A. (2011). Effect sizes and statistical methods for meta-analysis in higher education. *Research in Higher Education*. doi: 10.1007/s11162-011-9232-5

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement*, *31*, 37–50. doi: 10.1111/j.1745-3984.1994.tb00433.x

*Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT I: Reasoning Test.* Research Report No. 2000-1, College Examination Board, New York, NY. Retrieved from `https://www.ets.org/Media/Research/pdf/RR-00-01-Bridgeman.pdf`

*Bridgeman, B., Pollack, J., & Burton, N. (2008). *Predicting grades in different types of college courses.* Research Report No. 2008-1, College Board, New York, NY.

*Bridgeman, B., & Wendler, C. (1991). Gender differences in predictors of college mathematics performance and in college mathematics course grades. *Journal of Educational Psychology*, *83*, 275–284. doi: 10.1037/0022-0663.83.2.275

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Buchmann, C., DiPrete, T. A., & McDaniel, A. (2008). Gender inequalities in education. *Annual Review of Sociology*, *34*, 319–37. doi: 10.1146/annurev.soc.34.040507.134719

Bundesagentur für Arbeit. (2004a). *Studienfeldbezogene Beratungstestserie Naturwissenschaften*. Nürnberg, Germany.

Bundesagentur für Arbeit. (2004b). *Studienfeldbezogene Beratungstestserie Wirtschaftswissenschaften*. Nürnberg, Germany.

*Burton, N. W., & Wang, M. (2005). *Predicting long-term success in graduate school: A collaborative validity study.* GRE Research Report No. 99-14R, Educational Testing Service, Princeton, NJ. Retrieved from https://www.ets.org/Media/Research/pdf/RR-05-03.pdf

*Calkins, D. S., & Whitworth, R. (1974). *Differential prediction of freshmen grade point average for sex and two ethnic classifications at a southwestern university.* Retrieved from ERIC database (ED102199). Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED102199

Camara, W. J. (2003). *Scoring the essay on the SAT writing section.* Research Summary RS-10, College Board, New York, NY. Retrieved from http://research.collegeboard.org/sites/default/files/publications/2012/7/researchinreview-2003-10-scoring-sat-essay-writing.pdf

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

*Casserly, P. L. (1982). *Older students and the SAT.* College Board Report No. 82-2, College Board, New York, NY. Retrieved from http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-1982-8-older-students-sat.pdf

Ceci, S. J., Williams, W. M., & Barnett, S. M. (2009). Women's underrepresentation in science: Sociocultural and biological considerations. *Psychological Bulletin*, *135*, 218–261. doi: 10.1037/a0014412

Chapman, D. S., Uggerslev, K. L., Carroll, S. A., Piasentin, K. A., & Jones, D. A. (2005). Applicant attraction to organizations and job choice: A

meta-analytic review of the correlates of recruiting outcomes. *Journal of Applied Psychology*, *90*, 928–944. doi: 10.1037/0021-9010.90.5.928

*Chou, T., & Huberty, C. J. (1990). *A freshman admissions prediction equation: An evaluation and recommendation.* Retrieved from ERIC database (ED333081). Retrieved from `http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED333018`

Clark, A. E. (1997). Job satisfaction and gender: Why are women so happy at work? *Labour Economics*, *4*, 341–372. doi: 10.1080/13504850802297855

*Clark, M. J., & Grandy, J. (1984). *Sex differences in the academic performance of Scholastic Aptitude Test takers.* College Board Report No. 84-8, College Board, New York, NY. Retrieved from `http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-1984-8-sex-differences-performance-sat.pdf`

Cleary, T. A. (1968). Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, *5*, 115–124. doi: 10.1111/j.1745-3984.1968.tb00613.x

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312. doi: 10.1037/0003-066X.45.12.1304

College Entrance Examination Board. (2004). *The SAT program handbook 2004–2005.* New York, NY: Author.

Conger, D., & Long, M. C. (2010). Why are men falling behind? Gender gaps in college performance and persistence. *Annals of the American Academy of Political and Social Science*, *627*, 184–214. doi: 10.1177/0002716209348751

Costa, P. T., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences*, *13*, 653–665. doi: 10.1016/0191-8869(92)90236-I

*Cowen, S., & Fiori, S. J. (1991). *Appropriateness of the SAT in selecting students for admission to California State University, Hay-*

ward. Paper presented at the annual meeting of the California Educational Research Association, San Diego, CA. Retrieved from ERIC database (ED343934). Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED343934

Coyle, T. R., & Pillow, D. R. (2008). SAT and ACT predict college GPA after removing *g*. *Intelligence*, *36*, 719–729. doi: 10.1016/j.intell.2008.05.001

*Crawford, P. L., Alferink, D. M., & Spencer, J. L. (1986). *Postdictions of college GPAs from ACT composite scores and high school GPAs: Comparisons by race and gender.* Retrieved from ERIC database (ED326541). Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED326541

Curley, W. E., & Schmitt, A. P. (1993). *Revising SAT-Verbal items to eliminate differential item functioning.* Report No. 93-2, College Examination Board, New York, NY. Retrieved from http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-1993-2-sat-verbal-items-eliminate-differential-item-functioning.pdf

Darlington, R. B. (1971). Another look at "cultural fairness". *Journal of Educational Measurement*, *8*, 71–82. doi: 10.1111/j.1745-3984.1971.tb00908.x

de Ayala, R. J. (2009). *The theory and practice of item-response theory.* New York, NY: Guilford Press.

De Raad, B., & Schouwenburg, H. C. (1996). Personality in learning and education: A review. *European Journal of Personality*, *10*, 303–336. doi: 10.1002/(SICI)1099-0984(199612)10:5⟨303::AID-PER262⟩3.3.CO;2-U

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*, 13–21. doi: 10.1016/j.intell.2006.02.001

Dlugosch, S. (2005). *Prognose von Studienerfolg dargestellt am Beispiel des Auswahlverfahrens der Bucerius Law School.* Aachen, Germany: Shaker.

Dollinger, S. J. (2011). "Standardized minds" or individuality? Admissions tests and creativity revisited. *Psychology of Aesthetics, Creativity, and the Arts*, *5*, 329–341. doi: 10.1037/a0023659

Donnellan, M. B., & Lucas, R. E. (2008). Age differences in the Big Five across the life span: Evidence from two national samples. *Psychology and Aging*, *23*, 558–566. doi: 10.1037/a0012897

Donnon, T., Paolucci, E. O., & Violato, C. (2007). The predictive validity of the MCAT for medical school performance and medical board licensing examinations: A meta-analysis of the published research. *Academic Medicine*, *82*, 100–106. doi: 10.1097/01.ACM.0000249878.25186.b7

Dougherty, C. (2007). *Introduction to econometrics* (3rd ed.). Oxford, England: Oxford University Press.

Duckworth, A. L., & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology*, *98*, 198–208. doi: 10.1037/0022-0663.98.1.198

Educational Testing Service. (2009). *ETS guidelines for fairness review of assessments*. Princeton, NJ: Author.

Eid, M., Gollwitzer, M., & Schmitt, M. (2010). *Statistik und Forschungsmethoden*. Weinheim, Germany: Beltz.

*Elliott, R., & Strenta, A. C. (1988). Effects of improving the reliability of the GPA on prediction generally and on comparative predictions for gender and race particularly. *Journal of Educational Measurement*, *25*, 333–347. doi: 10.1111/j.1745-3984.1988.tb00312.x

Ellis, L., Karadi, K., Hershberger, S., Field, E., Wersinger, S., Pellis, S., ... Hetsroni, A. (2008). *Sex differences: Summarizing more than a century of scientific research*. New York, NY: Psychology Press.

Else-Quest, N., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*, 103–127. doi: 10.1037/a0018053

Fietze, S., Holst, E., & Tobsch, V. (2010). *Germany's Next Top Manager: Does personality explain the gender career gap?* Discussion Paper No. 5110, Bonn, Germany.

Fischer, F., Schult, J., & Hell, B. (2012). Sex differences in secondary school success: Why female students perform better. *European Journal of Psychology of Education*. doi: 10.1007/s10212-012-0127-4

Fischer, F., Schult, J., & Hell, B. (in revision). Unterschätzen deutsche Studierfähigkeitstests die Studienleistungen von Frauen – und wenn ja, warum?

Fischer, F. T., Schult, J., & Hell, B. (2013). Sex-specific differential prediction of college admission tests: A meta-analysis. *Journal of Educational Psychology*, *105*, 478–488. doi: 10.1037/a0031956

Fischer, L., & Fischer, O. (2005). Arbeitszufriedenheit: Neue Stärken und alte Risiken eines zentralen Konzepts der Organisationspsychologie. *Wirtschaftspsychologie*(1), 5–20.

Formazin, M., Schroeders, U., Köller, O., Wilhelm, O., & Westmeyer, H. (2011). Studierendenauswahl im Fach Psychologie: Testentwicklung und Validitätsbefunde. *Psychologische Rundschau*, *62*, 221–236. doi: 10.1026/0033-3042/a000093

Fox, J. (1991). *Regression diagnostics.* Newbury Park, CA: Sage. doi: 10.4135/9781412985604

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or *g*? The relationship between the scholastic assessment test and general cognitive ability. *Psychological Science*, *15*, 373–378. doi: 10.1111/j.0956-7976.2004.00687.x

Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 International Standard Classification of Occupations. *Social Science Research*, *25*, 201–239. doi: 10.1006/ssre.1996.0010

Geary, D. C. (2010). *Male, female: The evolution of human sex differences* (2nd ed.). Washington, DC: American Psychological Association.

Gebel, M., & Pfeiffer, F. (2010). Educational expansion and its heterogeneous returns for wage workers. *Schmollers Jahrbuch*, *130*, 19–42. doi: 10.3790/schm.130.1.19

Goldstein, D., Haldane, & Mitchell, C. (1990). Sex differences in visual-spatial ability: The role of performance factors. *Memory & Cognition*,

18, 546–550. doi: 10.3758/BF03198487

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley. doi: 10.1037/13240-000

Gulliksen, H., & Wilks, S. S. (1950). Regression tests for several samples. *Psychometrika*, *15*(2), 91–114. doi: 10.1007/BF02289195

Hahn, E., Gottschling, J., & Spinath, F. M. (2012). Short measurements of personality – Validity and reliability of the GSOEP Big Five Inventory (BFI-S). *Journal of Research in Personality*, *46*, 355–359. doi: 10 .1016/j.jrp.2012.03.008

Hamilton, L. C. (1992). *Regression with graphics*. Belmont, CA: Duxbury.

Hannon, B. (2012). Test anxiety and performance-avoidance goals explain gender differences in SAT-V, SAT-M, and overall SAT scores. *Personality and Individual Differences*, *53*, 816–820. doi: 10.1016/ j.paid.2012.06.003

Heine, C., Briedis, K., Didi, H.-J., Haase, K., & Trost, G. (2006). *Auswahl- und Eignungsfeststellungsverfahren beim Hochschulzugang in Deutschland und ausgewählten Ländern: Eine Bestandsaufnahme.* Kurzinformation A 3/2006, Hochschul-Informations-System, Hannover, Germany. Retrieved from `http://www.his.de/pdf/pub_kia/kia200603 .pdf`

Hell, B., Boramir, I., Schaar, H., & Schuler, H. (2006). Interne Personalauswahl und Personalentwicklung in deutschen Unternehmen. *Wirtschaftspsychologie*(1), 2–22.

Hell, B., Päßler, K., & Schuler, H. (2009). Was-studiere-ich.de: Konzept, Nutzen und Anwendungsmöglichkeiten. *Zeitschrift für Studium und Beratung*(4), 9–14.

Hell, B., Trapmann, S., & Schuler, H. (2007). Eine Metaanalyse der Validität von fachspezifischen Studierfähigkeitstests im deutschsprachigen Raum. *Empirische Pädagogik*, *21*, 251–270.

*Hewitt, B. N., & Goldman, R. D. (1975). Occam's razor slices through the myth that college women overachieve. *Journal of Educational Psychology*, *67*, 325–330. doi: 10.1037/h0077010

Hinz, T., & Gartner, H. (2012). Geschlechtsspezifische Lohnunterschiede

in Branchen, Berufen und Betrieben. *Zeitschrift für Soziologie*, *34*, 22–39.

*Hogrebe, M. C., Ervin, L., Dwinell, P. L., & Newman, I. (1983). The moderating effects of gender and race in predicting the academic performance of college developmental students. *Educational and Psychological Measurement*, *43*, 523–530. doi: 10.1177/001316448304300221

Holden, C. (1989). Court ruling rekindles controversy over SATs. *Science*, *243*, 885–887. doi: 10.1126/science.2919279

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

*House, J. D. (1988). *Gender differences in prediction of graduate course performance from admissions test scores: An empirical example of statistical methods for investigating prediction bias.* Paper presented at the annual forum of the Association for Institutional Research, Minneapolis, MN. Retrieved from ERIC database (ED424810). Retrieved from `http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED424810`

*House, J. D., & Keeley, E. J. (1993). *Differential prediction of graduate student achievement from Miller Analogies Test scores.* Paper presented at the annual meeting of the Illinois Association for Institutional Research, Oakbrook Terrace, IL. Retrieved from ERIC database (ED364605). Retrieved from `http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED364605`

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. doi: 10.1080/10705519909540118

Huff, K. L., Koenig, J. A., Treptau, M. M., & Sireci, S. G. (1999). Validity of MCAT scores for predicting clerkship performance of medical students grouped by sex and ethnicity. *Academic Medicine*, *74*, 41–44. doi: 10.1097/00001888-199910000-00035

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA:

Sage.

Irwing, P., & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta-analysis. *British Journal of Psychology*, *96*, 505–524. doi: 10.1348/000712605X53542

Jackson, D. N., & Rushton, J. P. (2006). Males have greater *g*: Sex differences in general mental ability from 100,000 17- to 18-year-olds on the Scholastic Assessment Test. *Intelligence*, *34*, 479–486. doi: 10.1016/j.intell.2006.03.005

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger/Greenwood.

Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychologische Rundschau*, *35*, 21–35.

Jäger, A. O., Süß, H., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test (BIS, Form 4)*. Göttingen, Germany: Hogrefe.

*Jones, R. F., & Vanyur, S. (1985). *An investigation of gender-related test bias for the Medical College Admission Test.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL. Retrieved from ERIC database (ED259024). Retrieved from `http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED259024`

Jöreskog, K. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*, 183–202. doi: 10.1007/BF02289343

Judge, T. A., Cable, D. M., Boudreau, J. W., & Bretz, Jr., R. D. (1995). An empirical investigation of the predictors of executive career success. *Personnel Psychology*, *48*, 485–519. doi: 10.1111/j.1744-6570.1995.tb01767.x

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology*, *87*, 530–541. doi: 10.1037//0021-9010.87.3.530

Judge, T. A., Higgins, C. A., Thoresen, C. J., & Barrick, M. R. (1999). The Big Five personality traits, general mental ability, and career success

across the life span. *Personnel Psychology*, *52*, 621–652. doi: 10.1111/j.1744-6570.1999.tb00174.x

Kadmon, G., Kirchner, A., Duelli, R., Resch, F., & Kadmon, M. (2012). Warum der Test für Medizinische Studiengänge (TMS)? *Zeitschrift für Evidenz, Fortbildung und Qualität im Gesundheitswesen*, *106*, 125–130. doi: 10.1016/j.zefq.2011.07.022

Kaiser, L. C. (2005). *Gender-job satisfaction differences across Europe – an indicator for labor market modernization.* Discussion Papers 537, Berlin, Germany.

Kalton, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, *2*, 303–314.

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, *19*, 81–97.

Kaufman, S. B. (2007). Sex differences in mental rotation and spatial visualization ability: Can they be accounted for by differences in working memory capacity? *Intelligence*, *35*, 211–223. doi: 10.1016/j.intell.2006.07.009

*Kirchner, G. L. (1993). Gender as a moderator variable in predicting success in a Master of Arts in Teaching program. *Educational and Psychological Measurement*, *53*, 155–157. doi: 10.1177/0013164493053001017

Kobrin, J. L., Deng, H., & Shaw, E. J. (2011). The association between SAT prompt characteristics, response features, and essay scores. *Assessing Writing*, *16*, 154–169. doi: 10.1016/j.asw.2011.01.001

Kobrin, J. L., Kim, Y., & Sackett, P. R. (2012). Modeling the predictive validity of SAT mathematics items using item characteristics. *Educational and Psychological Measurement*, *72*, 99–119. doi: 10.1177/0013164411404410

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT for predicting first-year college grade point average.* Research Report No. 2008-5, College Board, New York, NY. Retrieved from `http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2008-5-validity-sat`

-predicting-first-year-college-grade-point-average.pdf

Koenig, J. A., Sireci, S. G., & Wiley, A. (1998). Evaluating the predictive validity of MCAT scores across diverse applicant groups. *Academic Medicine*, *73*, 1095–1106. doi: 10.1097/00001888-199810000-00021

Koenig, K. A., Frey, M. C., & Detterman, D. K. (2008). ACT and general cognitive ability. *Intelligence*, *36*, 153–160. doi: 10.1016/j.intell.2007.03.005

Kroh, M. (2010). *Documentation of sample sizes and panel attrition in the German Socio Economic Panel (SOEP) (1984 until 2009).* Data Documentation 50, Berlin, Germany.

Kuncel, N. R., Credé, M., & Thomas, L. L. (2007). A meta-analysis of the predictive validity of the Graduate Management Admission test (GMAT) and undergraduate grade point average (UGPA) for graduate student academic performance. *Academy of Management Learning and Education*, *6*, 51–68. doi: 10.5465/AMLE.2007.24401702

Kuncel, N. R., & Hezlett, S. A. (2007). Standardized tests predict graduate students' success. *Science*, *315*, 1080–1081. doi: 10.1126/science.1136618

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. doi: 10.1037/Q033-2909.127.1.162

Kuncel, N. R., Wee, S., Serafin, L., & Hezlett, S. A. (2010). The validity of the Graduate Record Examination for master's and doctoral programs: A meta-analytic investigation. *Educational and Psychological Measurement*, *70*, 340–352. doi: 10.1177/0013164409344508

*Kyei-Blankson, L. (2005). *Predictive validity, differential validity, and differential prediction of the subtests of the Medical College Admission Test.* Doctoral dissertation. Retrieved from `http://rave.ohiolink.edu/etdc/view?acc_num=ohiou1125524238`

Lapierre, L. M., & Hackett, R. D. (2007). Trait conscientiousness, leader-member exchange, job satisfaction and organizational citizen-

ship behaviour: A test of an integrative model. *Journal of Occupational and Organizational Psychology*, *80*, 539–554. doi: 10.1348/096317906X154892

Lawrence, I. M., Curley, W. E., & McHale, F. J. (1988). *Differential item functioning for males and females on SAT-Verbal reading subscore items.* Report No. 88-4, College Examination Board, New York, NY. Retrieved from `http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-1988-4-differential-item-functioning-males-females-sat.pdf`

Lawshe, C. H. (1983). A simplified approach to the evaluation of fairness in employee selection procedures. *Personnel Psychology*, *36*, 601–608. doi: 10.1111/j.1744-6570.1983.tb02237.x

Leonard, D. K., & Jiang, J. (1999). Gender bias and the college predictions of the SATs: A cry of despair. *Research in Higher Education*, *40*, 375–407.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R* (2nd ed.). Göttingen, Germany: Hogrefe.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, *136*, 1123–1135. doi: 10.1037/a0021276

Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research*, *43*, 139–161. doi: 10.2307/1169933

Linn, R. L. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology*, *63*, 507–512. doi: 10.1037//0021-9010.63.4.507

Linn, R. L. (1982). Admissions testing on trial. *American Psychologist*, *37*, 279–291. doi: 10.1037/0003-066X.37.3.279

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement*, *21*, 33–47. doi: 10.1111/j.1745-3984.1984.tb00219.x

Linn, R. L. (1990a). Admissions testing: Recommended uses, validity, differential prediction, and coaching. *Applied Measurement in Education*, *3*, 297–318. doi: 10.1207/s15324818ame0304_1

Linn, R. L. (1990b). Comments on Atkinson and Geiser: Considerations for college admissions testing. *Educational Researcher*, *38*, 677–679. doi: 10.3102/0013189X09351982

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis.* Thousand Oakes, CA: Sage.

Lubinski, D. S., & Benbow, C. (2007). Sex differences in personal attributes for the development of scientific expertise. In S. J. Ceci & W. M. Williams (Eds.), *Why aren't more women in science: Top researchers debate the evidence* (pp. 79–100). Washington, DC: American Psychological Association. doi: 10.1037/11546-007

*Luthy, T. L. (1996). *Validity and prediction bias of grade performance from Graduate Record Examination scores for students at Northern Illinois University: Age and gender considerations.* Doctoral dissertation, available from ProQuest Dissertations and Theses database (UMI No. 9716551).

Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis. *Intelligence*, *32*, 481–498. doi: 10.1016/j.intell.2004.06.008

Lynn, R., & Kanazawa, S. (2011). A longitudinal study of sex differences in intelligence at ages 7, 11 and 16 years. *Personality and Individual Differences*, *51*, 321–324. doi: 10.1016/j.paid.2011.02.028

*Lynn, R., & Mau, W. (2001). Ethnic and sex differences in the predictive validity of the Scholastic Achievement Test for college grades. *Psychological Reports*, *88*, 1099–1104.

Magnusson, D. (1967). *Test theory.* Reading, MA: Addison-Wesley.

Mattern, K. D., & Patterson, B. F. (2013). Test of slope and intercept bias in college admissions: A response to Aguinis, Culpepper, and Pierce (2010). *Journal of Applied Psychology*, *98*, 134–147. doi: 10.1037/a0030610

Mattern, K. D., Patterson, B. F., Shaw, E. J., Kobrin, J. L., & Barbuti, S. M. (2008). *Differential validity and prediction of the SAT.* Research Report No. 2008-4, College Board, New York, NY. Retrieved from `http://professionals.collegeboard.com/`

data-reports-research/sat/validity-studies

Mattern, K. D., Shaw, E. J., & Kobrin, J. L. (2011). An alternative presentation of incremental validity: Discrepant SAT and HSGPA performance. *Educational and Psychological Measurement*, *71*, 638–662. doi: 10.1177/0013164410383563

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*, 1–10. doi: 10.1016/j.intell.2008.08.004

Meade, A. W., & Fetzer, M. (2009). Test bias, differential prediction, and a revised approach for determining the suitability of a predictor in a selection context. *Organizational Research Methods*, *12*, 738–761. doi: 10.1177/1094428109331487

Meade, A. W., & Tonidandel, S. (2010). Not seeing clearly with Cleary: What test bias analyses do and do not tell us. *Industrial and Organizational Psychology*, *3*, 192–205. doi: 10.1111/j.1754-9434.2010.01223.x

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, *30*, 577–605. doi: 10.1207/s15327906mbr3004_6

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*, 461–473. doi: 10.1007/S11336-007-9039-7

Nagl, W. (1992). *Statistische Datenanalyse mit SAS*. Frankfurt/Main, Germany: Campus Verlag.

*Nauels, H., & Meyer, M. (1997). Untersuchungen zur Vorhersagekraft des TMS: Differentielle Aspekte der Studienerfolgsprognose und Testfairneß. In G. Trost (Ed.), *Test für Medizinische Studiengänge (TMS): Studien zur Evaluation (21. Arbeitsbericht)* (pp. 76–134). Bonn, Germany: ITB.

Neisser, U., Boodoo, G., Thomas J. Bouchard, J., Boykin, A. W., Brody, N., Ceci, S. J., ... Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77–101. doi: 10.1037//0003-066X.51.2.77

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist*, *67*, 130–159. doi: 10.1037/

a0026699

Norborg, J. M. (1984). A warning regarding the simplified approach to the evaluation of test fairness in employee selection procedures. *Personnel Psychology*, *37*, 483–486. doi: 10.1111/j.1744-6570.1984.tb00524.x

Norris, D., Oppler, S., Kuang, D., Day, R., & Adams, K. (2006). *The College Board SAT Writing validation study: An assessment of predictive and incremental validity.* Research Report No. 2006-2, College Board, New York, NY. Retrieved from `http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2006-2-sat-writing-validation-predictive-incremental.pdf`

OECD. (2010). *PISA 2009 results: Overcoming social background – equity in learning opportunities and outcomes (volume II).* OECD Publishing. doi: 10.1787/9789264091504-en

Oh, I.-S., Schmidt, F. L., Shaffer, J. A., & Le, H. (2008). The Graduate Management Admission Test (GMAT) is even more valid than we thought: A new development in meta-analysis and its implications for the validity of the GMAT. *Academy of Management Learning & Education*, *7*, 563–570. doi: 10.5465/AMLE.2008.35882196

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Thousand Oaks, CA: Sage.

*Pape, T. E. (1992). *Selected predictors of examination for professional practice in psychology scores among graduates of Western Conservative Baptist Seminary's doctoral program in clinical psychology.* Doctoral dissertation, available from ProQuest Dissertations and Theses database (UMI No. 9302769).

Patterson, B. F., & Mattern, K. D. (2011). *Validity of the SAT for predicting first-year grades: 2008 SAT validity sample.* Statistical Report No. 2011-5, College Board, New York, NY. Retrieved from `http://professionals.collegeboard.com/profdownload/Validity_of_the_SAT_for_Predicting_First_Year_College_Grade_Point_Average.pdf`

*Patterson, B. F., Mattern, K. D., & Kobrin, J. L. (2009). *Validity of the SAT for predicting FYGPA: 2007 SAT validity sample.* Statis-

tical Report No. 2009-1, College Board, New York, NY. Retrieved from http://research.collegeboard.org/sites/default/files/publications/2012/7/statisticalreport-2009-1-validity-sat-1st-yr-gpa-2007-sample.pdf

Patterson, F., & Ferguson, E. (2010). Selection for medical education and training. In T. Swanwick (Ed.), *Understanding medical education: Evidence, theory and practice* (pp. 352–365). West Sussex, UK: Wiley. doi: 10.1002/9781444320282.ch24

*Patton, T. K. (1998). *Differential prediction of college performance between gender.* Unpublished Report. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED427030

*Pennock-Román, M. (1994). *College major and gender differences in the prediction of college grades.* College Board Report No. 94-2, College Board, New York, NY. Retrieved from http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-1994-2-college-major-gender-differences-prediction-college-grades.pdf

Poropat, A. E. (2009). A meta-analysis of the five-factor model of personality and academic performance. *Psychological Bulletin*, *135*, 322–338. doi: 10.1037/a0014996

*Qualls, A. L., & Ansley, T. N. (1995). The predictive relationship of ITBS and ITED to measures of academic success. *Educational and Psychological Measurement*, *55*, 485–498. doi: 10.1177/0013164495055003016

*Ramist, L., C. Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups.* Report No. 93-1, College Board, New York, NY. Retrieved from http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-1993-1-student-group-differences-predicting-college-grades.pdf

*Reuben, T. C. (2003). *Investigating test fairness of GRE scores for veterinary student selection.* Doctoral dissertation, available from ProQuest Dissertations and Theses database (UMI No. 3083156).

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of

university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, *138*, 353–387. doi: 10.1037/a0026838

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

Roth, P. L., BeVier, C. A., Switzer III, F. S., & Schippmann, J. S. (1996). Meta-analyzing the relationship between grades and job performance. *Journal of Applied Psychology*, *81*, 548–556. doi: 10.1037/0021-9010.81.5.548

Roth, P. L., & Clarke, R. L. (1998). Meta-analyzing the relation between grades and salary. *Journal of Vocational Behavior*, *53*, 386–400. doi: 10.1006/jvbe.1997.1621

Sackett, P. R. (2003). Stereotype threat in applied selection settings: A commentary. *Human Performance*, *16*, 295–309. doi: 10.1207/S15327043HUP1603_6

Sackett, P. R., Borneman, M. J., & Connelly, B. S. (2008). High-stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist*, *63*, 215–227. doi: 10.1037/0003-066X.63.4.215

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2005). On interpreting research on stereotype threat and test performance. *American Psychologist*, *60*, 271–272. doi: 10.1037/0003-066X.60.3.271

Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, *85*, 112–118. doi: 10.1037/0021-9010.85.1.112

Sanber, S. R., & Millman, J. (1987). *Gender and race effects on standardized tests predictive validity: A meta-analytical study.* Washington, DC: Paper presented at the Annual Meeting of the American Educational Research Association (April 20–24). Retrieved from `http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED286914`

Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation*

models (pp. 181–204). Newbury Park, CA: Sage.

Schmidt, F. L., & Hunter, J. E. (1982). Two pitfalls in assessing fairness of selection tests using the regression model. *Personnel Psychology*, *35*, 601–607. doi: 10.1111/j.1744-6570.1982.tb02212.x

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262–274. doi: 10.1037/0033-2909.124.2.262

Schnell, R. (1994). *Graphisch gestützte Datenanalyse*. München, Germany: Oldenbourg. Retrieved from `http://www.ub.uni-konstanz.de/kops/volltexte/2008/5613/`

Schnell, R., Hill, P. B., & Esser, E. (2005). *Methoden der empirischen Sozialforschung* (7th ed.). München, Germany: Oldenbourg.

Schuler, H., Funke, U., & Baron-Boldt, J. (1990). Predictive validity of school grades: A meta-analysis. *Applied Psychology: An International Review*, *39*, 89–103. doi: 10.1111/j.1464-0597.1990.tb01039.x

Schuler, H., Hell, B., Trapmann, S., Schaar, H., & Boramir, I. (2007). Die Nutzung psychologischer Verfahren der externen Personalauswahl in deutschen Unternehmen: Ein Vergleich über 20 Jahre. *Zeitschrift für Personalpsychologie*, *6*, 60–70. doi: 10.1026/1617-6391.6.2.60

Schult, J. (2012). Prädiktoren des Berufserfolgs von Hochschulabsolventen: Befunde aus dem Sozio-Ökonomischen Panel. *Wirtschaftspsychologie*, *14*, 82–91.

Schult, J., Fischer, F., & Hell, B. (2010). *Differenzielle Validität und differenzielle Prognose bei Auswahlverfahren: Konventionen und Perspektiven.* Poster presented at the 47th Kongress der Deutschen Gesellschaft für Psychologie, September 26–30, 2010, Bremen, Germany. Retrieved from `http://www.psychologie.uni-konstanz.de/index.php?eID=tx_nawsecuredl&file=fileadmin/psychologie/genderfairness/Schult_et_al_DGPs_2010.pdf`

Schult, J., Hell, B., Päßler, K., & Schuler, H. (2013). Sex-specific differential prediction of academic achievement by German ability tests. *International Journal of Selection and Assessment*, *21*, 130–134. doi:

10.1111/ijsa.12023

Schupp, J. (2009). 25 Jahre Sozio-oekonomisches Panel – Ein Infrastrukturprojekt der empirischen Sozial- und Wirtschaftsforschung in Deutschland. *Zeitschrift für Soziologie*, *38*, 350–357.

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (2nd ed., pp. 257–278). New York, NY: Russell Sage Foundation.

Shaw, E. J., Kobrin, J. L., Patterson, B. F., & Mattern, K. D. (2012). *The validity of the SAT for predicting cumulative grade point average by college major.* Research Report No. 2012-6, College Examination Board, New York, NY. Retrieved from `http://research.collegeboard.org/sites/default/files/publications/2012/8/researchreport-2012-6-validity-sat-predicting-cumulative-gpa-major.pdf`

*Siegert, K. O. (2007). *Predicting success in graduate management doctoral programs.* GMAC Research Reports No. RR-07-10, Graduate Management Admission Council, McLean, VA. Retrieved from `http://www.gmac.com/~/media/Files/gmac/Research/validity-and-testing/RR0710_DoctoralValidity.pdf`

*Sireci, S. G., & Talento-Miller, E. (2006). Evaluating the predictive validity of Graduate Management Admission Test scores. *Educational and Psychological Measurement*, *66*, 305–317. doi: 10.1177/0013164405282455

Snyder, T. D., & Dillow, S. A. (2012). *Digest of education statistics 2011 (NCES 2012-001).* Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U. S. Department of Education. Retrieved from `http://nces.ed.gov/pubs2012/2012001.pdf`

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Spearman, C. (1904). General intelligence. *The American Journal of Psychology*, *15*, 201–292. doi: 10.2307/1412107

Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science? *American Psychologist*, *60*, 950–958. doi: 10.1037/

0003-066X.60.9.950

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology*, *35*, 4–28. doi: 10.1006/jesp.1998.1373

Statistisches Bundesamt. (2012). *Frauenanteile der Studierenden, Absolventen und des Personals an Hochschulen.* Wiesbaden, Germany. Retrieved from `https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/BildungForschungKultur/Hochschulen/Tabellen/FrauenanteileAkademischeLaufbahn.html`

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*, 613–629. doi: 10.1037//0003-066X.52.6.613

Steinmayr, R., Beauducel, A., & Spinath, B. (2010). Do sex differences in a faceted model of fluid and crystallized intelligence depend on the method applied? *Intelligence*, *38*, 101–110. doi: 10.1016/j.intell.2009.08.001

Stemler, S. E. (2012). What should university admissions tests predict? *Educational Psychologist*, *47*, 5–17. doi: 10.1080/00461520.2011.611444

Sternberg, R. J., Bonney, C. R., Gabora, L., & Merrifield, M. (2012). WICS: A model for college and university admissions. *Educational Psychologist*, *47*, 30–41. doi: 10.1080/00461520.2011.638882

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99–110). Chichester, England: Wiley.

*Stricker, L. J., Rock, D. A., & Burton, N. W. (1993). Sex differences in predictions of college grades from Scholastic Aptitude Test scores. *Journal of Educational Psychology*, *85*, 710–718. doi: 10.1037/0022-0663.85.4.710

Su, R., Rounds, J., & Armstrong, P. I. (2009). Men and things, women and people: A meta-analysis of sex differences in interests. *Psychological Bulletin*, *135*, 859–884. doi: 10.1037/a0017364

*Swinton, S. S. (1987). *The predictive validity of the restructured GRE with particular attention to older students.* GRE Board Report No. 83-25P, Educational Testing Service, Princeton, NJ. Retrieved from `https://www.ets.org/Media/Research/pdf/RR-87-22-Swinton.pdf`

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Pearson.

Tacq, J. (1997). *Multivariate analysis techniques in social science research.* London, England: Sage.

*Talento-Miller, E. (2008). Generalizability of GMAT validity to programs outside the U. S. *International Journal of Testing*, *8*, 127–142. doi: 10.1080/15305050802001193

*Talento-Miller, E. (2009). *Validity study of non-MBA programs.* GMAC Research Reports No. RR-09-11, Graduate Management Admission Council, McLean, VA.

*Thomas, C. L. (1973). *The overprediction phenomenon among black collegians: Some preliminary considerations.* Paper presented at the annual meeting of American Educational Research Association, New Orleans, LA. Retrieved from ERIC database (ED076679). Retrieved from `http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED076679`

*Thomas, C. L. (1979). Relative effectiveness of high school grades for predicting college grades: Sex and ability level effects. *Journal of Negro Education*, *48*, 6–13. doi: 10.2307/2294611

Trapmann, S. (2008). *Mehrdimensionale Studienerfolgsprognose: Die Bedeutung kognitiver, temperamentsbedingter und motivationaler Prädiktoren für verschiedene Kriterien des Studienerfolgs.* Berlin, Germany: Logos.

Trapmann, S., Hell, B., Hirn, J. W., & Schuler, H. (2007). Meta-analysis of the relationship between the Big Five and academic success at university. *Zeitschrift für Psychologie / Journal of Psychology*, *215*, 132–151. doi: 10.1027/0044-3409.215.2.132

Trapmann, S., Hell, B., Weigand, S., & Schuler, H. (2007). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs – eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie*, *21*, 11-27. doi: 10.1024/1010

-0652.21.1.11

Trattner, M. H., & O'Leary, B. S. (1980). Sample sizes for specified statistical power in testing for differential validity. *Journal of Applied Psychology*, *65*, 127–134. doi: 10.1037/0021-9010.65.2.127

Trost, G. (2003). *Deutsche und internationale Studierfähigkeitstests. Arten, Brauchbarkeit, Handhabung.* Dokumente und Materialien, Band 51, DAAD, Bonn, Germany.

Trost, G. (2005). Studierendenauswahl durch die Hochschulen: Welche Schritte sind zu tun? *Psychologische Rundschau*, *56*, 140–142. doi: 10.1026/0033-3042.56.2.140

Trost, G., Nauels, H.-U., & Klieme, E. (1998). The relationship between different criteria for admission to medical school and student success. *Assessment in Education: Principles, Policy & Practice*, *5*, 247–254. doi: 10.1080/0969594980050206

von Stumm, S., Hell, B., & Chamorro-Premuzic, T. (2011). The hungry mind intellectual curiosity is the third pillar of academic performance. *Perspectives on Psychological Science*, *6*, 574–588. doi: 10.1177/1745691611421204

Wainer, H., Saka, T., & Donogue, J. R. (1992). The validity of the SAT at the University of Hawaii: A riddle wrapped in an enigma. *Educational Evaluation and Policy Analysis*, *15*, 91–98. doi: 10.2307/1164254

Walton, G. M., & Spencer, S. J. (2009). Latent ability: Grades and test scores systematically underestimate the intellectual ability of negatively stereotyped students. *Psychological Science*, *20*, 1132–1139. doi: 10 .1111/j.1467-9280.2009.02417.x

Waugh, C. K., & Gronlund, N. E. (2013). *Assessment of student achievement* (10th ed.). Boston, MA: Pearson.

Weaver, B., & Wuensch, K. L. (2013). SPSS and SAS programs for comparing Pearson correlations and OLS regression coefficients. *Behavior Research Methods*. doi: 10.3758/s13428-012-0289-7

Willingham, W. W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, *39*, 1–37. doi: 10.1111/j.1745-3984.2002.tb01133.x

*Wilson, K. M. (1982). *A study of the validity of the restructured GRE aptitude tests for predicting first-year performance in graduate study.* GRE Board Research Report No. 78-6R, Educational Testing Service, Princeton, NJ. Retrieved from `https://www.ets.org/Media/Research/pdf/RR-82-34-Wilson.pdf`

Wolniak, G. C., & Pascarella, E. T. (2005). The effects of college major and job field congruence on job satisfaction. *Journal of Vocational Behavior*, *67*, 233–251. doi: 10.1016/j.jvb.2004.08.010

Wood, J. A. (2008). Methodology for dealing with duplicate study effects in a meta-analysis. *Organizational Research Methods*, *11*, 79–95. doi: 10.1177/1094428106296638

Wooldridge, J. M. (2006). *Introductory econometrics: A modern approach* (3rd ed.). Mason, OH: Thomson.

Wüstenberg, S., Greiff, S., Molnar, G., & Funke, J. (submitted). Determinants of cross-national gender differences in complex problem solving competency.

*Wynne, W. D. (2003). *An investigation of ethnic and gender intercept bias in the SAT's prediction of college freshman academic performance.* Doctoral dissertation, available from ProQuest Dissertations and Theses database (UMI No. 3116464).

*Young, J. W. (1994). Differential prediction of college grades by gender and by ethnicity: A replication study. *Educational and Psychological Measurement*, *54*, 1022–1029. doi: 10.1177/0013164494054004019

Young, J. W., & Kobrin, J. L. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis.* College Board No. 2001-6, College Board, New York, NY. Retrieved from `http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2001-6-differential-validity-prediction-college-admission-testing-review.pdf`

*Zeidner, M. (1987). A cross-cultural test of sex bias in the predictive validity of scholastic aptitude examinations: Some Israeli findings. *Evaluation and Program Planning*, *10*, 289–295. doi: 10.1016/0149-7189(87)90041-3

Zhang, Y., Dorans, N. J., & Matthews-López, J. L. (2005). *Using DIF dissection method to assess effects of item deletion.* College Board No. 2005-10, College Board, New York, NY. Retrieved from `http://research.collegeboard.org/sites/default/files/publications/2012/7/researchreport-2005-10-using-dif-dissection-method-assess-effects-item-deletion.pdf`

Zwick, R. (2002). *Fair game? The use of standardized admissions tests in higher education*. New York, NY: Routledge.

Zwick, R. (2007). *College admission testing.* National Association for College Admission Counseling. Retrieved from `http://stage.nacacnet.org/research/PublicationsResources/Marketplace/Documents/TestingWhitePaper.pdf`

Zwick, R., Sklar, J. C., Wakefield, G., Hamilton, C., Norman, A., & Folsom, D. (2008). Instructional tools in educational measurement and statistics (ITEMS) for school personnel: Evaluation of three web-based training modules. *Educational Measurement: Issues and Practice*, *27*(2), 14–27. doi: 10.1111/j.1745-3992.2008.00119.x

A Supporting Online Material

Supporting Online Material for *Sex-Specific Differential Prediction of College Admission Tests: A Meta-Analysis*

Table A.1: Overview of the Studies Included in the Meta-Analysis of Residuals: Average Effect Sizes, Predictor and Criterion Information by Sample

					Predictor(s)			
					Admission test		Admission test and HGPA/UGPA	
Reference	Sample description	Test name	Criterion	N	d_f	d_m	d_f	d_m
American College Testing Program, 1973	College A	ACT	FSGPA	1,703	.30	–.20	.25	–.16
	College B	ACT	FSGPA	1,281	.28	–.21	.21	–.17
	College C	ACT	FSGPA	593	.35	–.33	.28	–.26
	College D	ACT	FSGPA	724	.27	–.19	.26	–.17
	College E	ACT	FSGPA	426	.44	–.16	.37	–.13
	College F	ACT	FSGPA	616	.36	–.20	.27	–.14
	College G	ACT	FSGPA	1,035	.46	–.23	.38	–.19
	College H	ACT	FSGPA	1,349	.25	–.16	.17	–.12
	College I	ACT	FSGPA	1,451	.35	–.26	.29	–.21
	College J	ACT	FSGPA	1,490	.10	–.15	.05	–.08
	College K	ACT	FSGPA	577	.30	–.26	.23	–.18
	College L	ACT	FSGPA	1,578	.28	–.21	.20	–.13
	College M	ACT	FSGPA	4,149	.22	–.14	.11	–.07
	College N	ACT	FSGPA	1,057	.43	–.20	.31	–.15
	College O	ACT	FSGPA	1,374	.46	–.30	.42	–.25
	College P	ACT	FSGPA	699	.52	–.28	.51	–.25
	College Q	ACT	FSGPA	1,220	.30	–.16	.19	–.10
	College R	ACT	FSGPA	638	.30	–.22	.24	–.18
	College S	ACT	FSGPA	694	.32	–.10	.13	–.04
Bridgeman et al., 2000		SAT I	FGPA	46,916	.14	–.15	.10	–.12
Bridgeman et al., 2008		SAT (revised)	FGPA	110,468	-	-	.11	–.12
Burton & Wang, 2005	Biology graduate students	GRE	CGGPA	145	-	-	.11	–.12
	Chemistry graduate students	GRE	CGGPA	134	-	-	.21	–.09
	Education graduate students	GRE	CGGPA	699	-	-	.03	–.09
	English graduate students	GRE	CGGPA	170	-	-	–.07	.11
	Psychology graduate students	GRE	CGGPA	155	-	-	.09	–.15
Casserly, 1982	College A, B and C	SAT	FGPA	1,540	-	-	.13	–.20
Chou & Huberty, 1990		SAT	Cumulative GPA after 9 months	3,378	-	-	.05	–.07
Clark & Grandy, 1984	Engineering students	SAT	FGPA	334	-	-	-	.00
	Science students	SAT	FGPA	296	-	-	.00	.00
	Business students	SAT	FGPA	437	-	-	.00	.00
Cowen & Fiori, 1991	Regular progressors	SAT	FGPA	642	-	-	–.02	.09
	Slower progressors	SAT	FGPA	181	-	-	–.05	.04
Elliot & Strenta, 1988		SAT and Achievement test	Cumulative raw GPA after 4 years	913	-	-	.10	–.07

Table A.1 (continued)

					Predictor(s)			
					Admission test		Admission test and HGPA/UGPA	
Reference	Sample description	Test name	Criterion	*N*	d_f	d_m	d_f	d_m
House, 1998	Psychology graduate students	GRE	Grade in Psychodiagnostics I	269	–.14	.34	-	-
House & Keeley, 1993		MAT	CGGPA	1,438	–.06	.43	-	-
Kyei-Blankson, 2005	School 1	MCAT	FGPA	209	.03	–.01	-	-
	School 2	MCAT	FGPA	136	–.09	.06	-	-
	School 3	MCAT	FGPA	193	.02	–.01	-	-
	School 4	MCAT	FGPA	520	.06	–.04	-	-
	School 5	MCAT	FGPA	291	–.07	.05	-	-
	School 6	MCAT	FGPA	372	–.05	.02	-	-
	School 7	MCAT	FGPA	262	–.09	.08	-	-
	School 8	MCAT	FGPA	256	.09	–.06	-	-
	School 9	MCAT	FGPA	226	.06	–.04	-	-
	School 10	MCAT	FGPA	188	.13	–.15	-	-
	School 11	MCAT	FGPA	173	.11	–.10	-	-
	School 12	MCAT	FGPA	132	–.10	.10	-	-
	School 13	MCAT	FGPA	140	.10	–.06	-	-
	School 14	MCAT	FGPA	89	.09	–.11	-	-
Luthy, 1996	Adult Continuing Education majors	GRE	CGGPA	388	.05	–.09	-	-
	Educational Administration majors	GRE	CGGPA	615	.16	–.15	-	-
	Engineering majors	GRE	CGGPA	376	.05	–.01	-	-
	Computer Science majors	GRE	CGGPA	298	.22	–.10	-	-
	English majors	GRE	CGGPA	367	.04	–.07	-	-
	Political Science majors	GRE	CGGPA	357	.01	.00	-	-
	Psychology majors	GRE	CGGPA	219	.17	–.31	-	-
	Communicative Disorders majors	GRE	CGGPA	229	.01	–.21	-	-
Lynn & Mau, 2001		SAT	Baccalaureate degree	3,512	.24	–.27	-	-
Pape, 1992		GRE	Examination for Professional Practice in Psychology	67	–.11	.04	-	-
Patterson et al., 2009		SAT (revised)	FGPA	159,286	.14	–.17	.10	–.12
Ramist et al., 1994		SAT	FGPA	46,379	.14	–.15	.10	–.10
Reuben, 2003		GRE	First year veterinary school GPA	634	–.02	.06	–.04	.10
Siegert, 2007		GMAT	FGPA	518	-	-	.09	–.07
Sireci & Talento-Miller, 2006		GMAT	FGPA	4,172	-	-	.00	.00
Stricker et al., 1993		SAT	FSGPA (unadjusted)	4,351	.12	–.14	.07	–.08
Swinton,1987	All subjects	GRE	FGPA	2,018	-	-	.01	–.03

Table A.1 (continued)

					Predictor(s)			
					Admission test		Admission test and HGPA/UGPA	
Reference	Sample description	Test name	Criterion	*N*	d_f	d_m	d_f	d_m
Talento-Miller, 2008		GMAT	Graduate program GPA	1,063	.15	–.06	-	-
Talento-Miller, 2009		GMAT	Mid-program GPA	1,333	-	-	.04	–.04
Thomas, 1973	College A	ACT	FGPA	415	-	-	.03	–.04
	College B	ACT	FGPA	2,727	-	-	.13	–.11
	College C	ACT	FGPA	1,203	-	-	.20	–.17
	College D	ACT	FGPA	861	-	-	.14	–.16
	College E	ACT	FGPA	1,692	-	-	.02	–.02
	College F	ACT	FGPA	1,404	-	-	.18	–.25
	College G	ACT	FGPA	1,980	-	-	.05	–.06
	College H	ACT	FGPA	1,726	-	-	.14	–.12
	College I	ACT	FGPA	868	-	-	.06	–.08
Wilson, 1982	All verbal subjects	GRE	First year GGPA	697	.03	–.04	-	-
	All quantitative subjects	GRE	First year GGPA	622	–.05	.01	-	-
Young, 1994		SAT	Cumulative college GPA	3,703	-	-	.12	–.15
Zeidner, 1987	Jewish subsample	SAT Hebrew version	FGPA	824	.16	–.25	-	-
	Arab subsample	SAT Arabic version	FGPA	364	.05	–.02	-	-

Note. Positive effect sizes indicate underprediction, negative effect sizes indicate overprediction. Dash indicates that the data is not provided for the sample or could not be calculated. d_f = standardized effect size for females; d_m = standardized effect size for males; ACT = American College Test; GRE = Graduate Record Examination; MAT = Miller Analogies Test; MCAT = Medical College Admission Test; GMAT = Graduate Management Admission Test; FSGPA = first semester GPA; FGPA = first year GPA; GGPA = graduate GPA; CGGPA = cumulative graduate GPA.

Table A.2: Overview of the Studies Included in the Summary of Differences in Regression Equations: Intercept/Slope Differences, Predictor and Criterion Information by Sample

					Significant intercept and/or slope differences	
Reference	Sample description	Test name	Criterion	*N*	Admission test as predictor	Admission test and HGPA/UGPA as predictor
Bridgeman & Wendler, 1991	College 1a	SAT-M	Calculus grades	1,050	-	no
	College 1b	SAT-M	Calculus grades	2,293	-	i
	College 2a	SAT-M	Calculus grades	106	-	i
	College 2b	SAT-M	Calculus grades	214	-	no
	College 3	SAT-M	Algebra grades	272	-	no
	College 4	SAT-M	Calculus grades	184	-	no
	College 5	SAT-M	Precalculus grades	183	-	no
	College 6	SAT-M	Calculus grades	129	-	i
Calkins & Whitworth, 1974		SAT	GPA attained for the first 30 or less college hours	3,237	-	i
Crawford et al., 1986	College G	ACT	College GPA	1,121	no	no
Dlugosch, 2005	2000 sample	Test of the BLS	GPA after 2 years	63	i	-
	2001 sample	Test of the BLS	GPA after 2 years	91	no	-
Hewitt & Goldman, 1975	Los Angeles campus	SAT	GPA	-	i	-
	Davis campus	SAT	GPA	-	i	-
	Irvine campus	SAT	GPA	-	i	-
	San Diego campus	SAT	GPA	-	i	-
Hogrebe et al., 1983		SAT	FGPA	345	-	i
Jones & Vanyur, 1985	School A	MCAT	FGPA	252	-	no
	School B	MCAT	FGPA	357	-	no
Kirchner, 1993		GRE	Graduate GPA from three consecutive terms	103	-	no
Nauels & Meyer, 1997	Human medicine students	TMS	Examination after 2 years	19,561	i	i, s
	Veterinary medicine students	TMS	Examination after 1 year	2,391	i, s	s
	Dentistry studenrs	TMS	Examination after 1 year	5,221	i	i
Patton, 1998	Biology students	ACT	Cumulative college GPA	195	no	-
	English students	ACT	Cumulative college GPA	254	i	-
	Finance students	ACT	Cumulative college GPA	257	no	-

Table A.2 (continued)

Reference	Sample description	Test name	Criterion	*N*	Significant intercept and/or slope differences: Admission test as predictor	Significant intercept and/or slope differences: Admission test and HGPA/UGPA as predictor
	Math students	ACT	Cumulative college GPA	60	no	-
	Psychology students	ACT	Cumulative college GPA	430	i	-
Pennock-Román, 1994	Non-Latino White Texas	SAT	FGPA	4,148	-	i
	Non-Latino White Massachusetts	SAT	FGPA	4,428	-	no
	Non-Latino White California: Public	SAT	FGPA	1,272	-	no
	Non-Latino White California: Private	SAT	FGPA	890	-	s
	African American Texas	SAT	FGPA	264	-	no
	African American Massachusetts	SAT	FGPA	178	-	no
	African American California: Private	SAT	FGPA	116	-	i
	Latino American Texas	SAT	FGPA	577	-	no
	Latino American Massachusetts	SAT	FGPA	116	-	no
	Latino American California: Private	SAT	FGPA	106	-	no
	Asian American Texas	SAT	FGPA	237	-	s
	Asian American Massachusetts	SAT	FGPA	301	-	no
	Asian American California: Public	SAT	FGPA	561	-	no
	Asian American California: Private	SAT	FGPA	107	-	s
Qualls & Ansley, 1995		ACT	FGPA	1,038	i	-
Thomas, 1979	Curriculum A	ACT	FSGPA	88	no	no
	Curriculum B	ACT	FSGPA	96	i	i
	Curriculum C	ACT	FSGPA	96	i	i
Wynne, 2003		SAT	GPA	836	i	i

Note. Dash indicates that the data is not provided for the sample or could not be calculated. i = significant intercept differences; s = significant slope differences; TMS = Test for medical study programs in Germany; SAT-M = Mathematics section of the SAT; ACT = American College Test; Test of the BLS = Admission test of the German Bucerius Law School; MCAT = Medical College Admission Test; GRE = Graduate Record Examination; FSGPA = first semester GPA; FGPA = first year GPA.