Fairness von Studierfähigkeitstests:

Ausmaß und Gründe für die Unterschätzung der Studienleistungen von Frauen

Dissertation

zur Erlangung des akademischen Grades

des Doktors der Naturwissenschaften

vorgelegt von

Franziska Fischer

an der Universität Konstanz

Mathematisch-Naturwissenschaftliche Sektion

Fachbereich Psychologie

Tag der mündlichen Prüfung: 01.02.2013

1. Referent: Prof. Dr. Benedikt Hell

2. Referent: Prof. Dr. Sabine Hochholdinger

Vorveröffentlichungen der Dissertation

Teilergebnisse dieser Dissertation wurden bereits in folgenden Beiträgen vorgestellt:

Publikationen

Fischer, F., Schult, J., & Hell, B. (2013). Sex-specific differential prediction of college admission tests: A meta-analysis. *Journal of Educational Psychology*, *105*(2), 478-488. doi: 10.1037/a0031956

Fischer, F., Schult, J. & Hell, B. (2013). Sex differences in secondary school success: Why female students perform better. *European Journal of Psychology of Education*, *28*(2), 529-543. doi: 10.1007/s10212-012-0127-4

Fischer, F., Schult, J. & Hell, B. (2012). *Unterschätzung der Studienleistungen von Frauen durch Studierfähigkeitstests: Erklärbar durch Persönlichkeits-eigenschaften?* Manuskript eingereicht zur Publikation.

Konferenzbeiträge

Fischer, F., Schult, J. & Hell, B. (2010). *Gender-Fairness von Studierfähigkeitstests: eine Metaanalyse zur differenziellen Prognose von Studienleistungen*. Präsentation auf dem 47. Kongress der Deutschen Gesellschaft für Psychologie, Bremen.

Fischer, F., Schult, J. & Hell, B. (2012). *Moderators of Sex-Specific Differential Prediction in College Admission Testing*. Presentation on the 30th International Congress of Psychology. Cape Town, South Africa.

Inhaltsverzeichnis

Zusammenfassung

In attraktiven Studiengängen übersteigen die Bewerberzahlen die zur Verfügung stehenden Studienplätze bei Weitem. Die Hochschulen müssen immer häufiger entscheiden, welche Studienbewerber sie aufnehmen wollen. Im Rahmen hochschuleigener Auswahlverfahren setzen deutsche Hochschulen neben der Abiturnote zunehmend Studierfähigkeitstests ein, welche sich in anderen Ländern seit Jahren etabliert haben. Die Tests haben damit einen Einfluss auf den Berufsweg von zahlreichen Studieninteressierten, weshalb hohe Fairnessanforderungen an die Tests gestellt werden. Amerikanische Studien liefern jedoch Hinweise, dass Studierfähigkeitstests die Studienleistungen von Frauen unterschätzen (d. h. eine geschlechtsspezifische differenzielle Prognose aufweisen; Young & Kobrin, 2001). Diese Annahme wird in der vorliegenden Arbeit auf ihre Generalisierbarkeit überprüft (Frage I). Zusätzlich wird untersucht, inwiefern sich die geschlechtsspezifische differenzielle Prognose auch für ausgewählte deutschsprachige Studierfähigkeitstests zeigt (Frage II), ob die Mitberücksichtigung der Abiturnote die Vorhersage-Fairness günstig beeinflusst und wenn ja, woran dies liegt (Frage III) und wie eine solche geschlechtsspezifische Unterschätzung des Studienerfolgs erklärt werden kann (Frage IV). Die Beantwortung der Forschungsfragen erfolgt anhand dreier Studien:

In der ersten Studie (Kapitel 3) wird der aktuelle Forschungsstand zur geschlechtsspezifischen differenziellen Prognose von Studienleistungen durch Studierfähigkeitstests (sowie durch Test und Abiturnote gemeinsam) in einer Metaanalyse zusammengefasst. Ferner werden Variablen identifiziert, die das Auftreten von differenzieller Prognose moderieren. Aufbauend auf den Ergebnissen der Metaanalyse wird in einer zweiten Studie untersucht, welche Persönlichkeitsmerkmale neben den kognitiven Fähigkeiten den Abiturerfolg von Frauen und Männern determinieren und damit zu einer fairen Vorhersage des Studienerfolgs beitragen (Kapitel 4). In der dritten und letzten Studie wird an einer großen Erstsemester-Stichprobe längsschnittlich die geschlechtsspezifische differenzielle Prognose von zwei deutschsprachigen Studierfähigkeitstests analysiert.

Außerdem wird getestet, ob die differenzielle Prognose alle Leistungsbereiche gleich stark betrifft und inwieweit Persönlichkeitsunterschiede die geschlechtsspezifische differenzielle Prognose erklären (Kapitel 5).

Zusammenfassend liefern die durchgeführten Studien folgende Antworten auf die formulierten Forschungsfragen.

Zu Frage I: Die Metaanalyse bestätigt eine generalisierbare Unterschätzung der Studienleistungen von Frauen durch Studierfähigkeitstests im Umfang von d = .14. Zusätzlich kann in der dritten Studie zum ersten Mal gezeigt werden, dass die geschlechtsspezifische differenzielle Prognose im oberen Leistungsbereich besonders stark ausfällt, d. h. Frauen werden insbesondere bei strengen Selektionsquoten durch Studierfähigkeitstests benachteiligt.

Zu Frage II: Die dritte Studie zeigt, dass auch die eingesetzten deutschen Studierfähigkeitstests die Studienleistungen von Frauen unterschätzen. Das Ausmaß der Unterschätzung liegt hierbei im Bereich der Ergebnisse der Metaanalyse.

Zu Frage III: Dass die Vorhersage des Studienerfolgs anhand eines Studierfähigkeitstests kombiniert mit der Abiturnote fairer ausfällt als durch einen Studierfähigkeitstest alleine, bestätigen sowohl die Ergebnisse der Metaanalyse als auch die Befunde der dritten Studie. Dies lässt vermuten, dass die Abiturnote neben den kognitiven Fähigkeiten auch Persönlichkeitsmerkmale wie Leistungsmotivation miterfasst, die entscheidend dafür sind, dass Frauen das vorhandene kognitive Potential besser in gute Studienleistungen umsetzen können. Die zweite Studie bestätigt diese Annahme.

Zu Frage IV: Weder das Testalter noch geschlechtsspezifische Mittelwerts-unterschiede im Test und in den Studiennoten können die Unterschätzung der Studienleistungen von Frauen durch Studierfähigkeitstests in der Metaanalyse vollständig aufklären, es zeigt sich jedoch für *Undergraduate-Tests* eine stärkere Unterschätzung der Studienleistungen von Frauen als für *Graduate-Tests*. Frauen und

Männer, die sich für ein *Graduate-Studium* entscheiden, scheinen homogener in Bezug auf ihre Persönlichkeitseigenschaften zu sein als *Undergraduate-Studenten*. Befunde aus Studie drei stützen diese Hypothese. Wenn die Persönlichkeitsmerkmale Selbstdisziplin oder Leistungsmotivation bei der Vorhersage des Studienerfolgs berücksichtigt werden kann sich die differenzielle Prognose von Studierfähigkeitstests verringern.

Für die Praxis machen die gewonnenen Erkenntnisse deutlich, dass der alleinige Einsatz von Studierfähigkeitstests besonders im Zusammenhang mit strengen Selektionsquoten problematisch ist und möglichst schon bereits bei der Testentwicklung bzw. bei der Konzeption des Auswahlverfahrens auf die Genderfairness geachtet werden sollte.

Summary

The number of university applicants exceeds the number of available places in attractive fields of study. German universities have started to use college admission tests in addition to high school grades as selection criterion, whereas in other countries these tests have already been used for years. Because of this major influence on the career of young people, the fairness standards are set very high for college admission tests. Existing studies show that American admission tests tend to underpredict the college performance of women (i.e. tests show sex-specific differential prediction; Young & Kobrin, 2001). The present dissertation explores whether this hypothesis can be generalized (question I). Further, it is investigated whether the sex-specific differential prediction results can be replicated for two German admission tests (question II), whether considering high school grades has a positive influence on the predictive fairness (question III), and how the sex-specific underprediction of college performance can be explained (question IV). To answer the research questions three studies were conducted:

In the first study (chapter 3) the current state of research concerning sex-specific differential prediction of college admission tests (as well as that of the combination of admission tests and high school grades) is summarized with the help of meta-analytic techniques. Moderators of sex-specific differential prediction are also identified. Building on the results of the meta-analysis the second study investigates, which personality traits determine the high school success of women after controlling for cognitive abilities (chapter 4). In the third and last study a large freshmen sample is followed longitudinally to examine the sex-specific differential prediction of two German admission tests. Additionally, it is tested whether sex-specific differential prediction is a problem of all performance ranges and whether personality traits can explain the differential prediction effect (chapter 5).

The three studies provide the following answers to the presented research questions:

Question I: The meta-analysis confirms a generalizable underprediction of women's academic performance by college admission tests (d = .14). Study 3 shows for the first time that the sex-specific differential prediction is especially a problem in the high performance range. In other words, women are disadvantaged by college admission tests particularly if the selection ratio is very strict.

Question II: Study 3 demonstrates that the applied German admission tests underpredict the college performance of women. The amount of underprediction is similar to the results of the meta-analysis.

Question III: Both the meta-analysis and study 3 illustrate that the prediction is fairer when admission tests are used in combination with high school grades. This suggests that high school grades account for personality traits, like achievement motivation, besides cognitive abilities. These personality traits seem to be crucial in order to transfer the cognitive potential in course achievement at college. Study 2 supports this conclusion.

Question IV: According to the meta-analysis, neither publication year of the tests nor sex-specific predictor or criterion differences can fully explain the underprediction of women's college performance. Yet, undergraduate college admission tests show more underprediction of women's academic performance than graduate admission tests. Male and female graduate students appear to be more similar concerning their personality traits. Results of study 3 support this hypothesis. Accounting for the personality traits self-discipline or achievement motivation can reduce partially the differential prediction of college admission tests.

From an applied point of view these results make clear that using admission tests as sole predictor leads to an unfair prediction, especially when only a small margin of applicants is selected. If possible, this problem should be already addressed during test development and during the development of the selection procedure respectively.

1 Einleitung

Wer erhält heutzutage einen Studienplatz in besonders attraktiven und zukunftsträchtigen Studiengängen wie beispielsweise in der Medizin? Diese Frage stellt sich immer häufiger in begehrten Studienfächern, da hier die Bewerberzahlen die zur Verfügung stehenden Studienplätze bei Weitem übersteigen. Die Universitäten haben die Qual der Wahl. Im Rahmen hochschuleigener Auswahlverfahren setzen deutsche Hochschulen neben der Abiturnote zunehmend fachspezifische Studierfähigkeitstests ein, welche sich in den Vereinigten Staaten von Amerika (USA) bereits seit Langem etabliert haben. Millionen von jungen Menschen absolvieren dort jährlich Aufnahmetests für Bachelor- und Master-Programme (Berry & Sackett, 2009). Studierfähigkeitstests kommt dementsprechend eine große Lenkungswirkung zu, da sie in vielen Fällen mitentscheiden, an welche Bewerber[1] die gefragten Studienplätze vergeben werden.

Neben der Auswahl von Studierenden spielen Studierfähigkeitstests bei der Beratung von Studieninteressierten eine große Rolle. Sowohl die Bundesagentur für Arbeit als auch Self-Assessment-Angebote im Internet setzen Studierfähigkeitstests im Beratungsalltag ein, so dass die Testergebnisse einen wichtigen Bestandteil für die Studien-, Berufs- und Laufbahnorientierung der Ratsuchenden bilden.

Wegen diesen weit reichenden Auswirkungen für den Berufs- und Karriereweg von Studieninteressenten werden hohe Anforderungen an Studierfähigkeitstests gestellt. Die Vorhersage der Tests muss fair und valide sein, um nicht zuletzt gegen rechtliche Klagen standhalten zu können (Trost, 2005). Die prognostische Validität von Studierfähigkeitstests wurde bereits hinreichend belegt (Hell, Trapmann & Schuler, 2007; Patterson & Mattern, 2011), wie *fair* diese Tests sind, ist hingegen besonders im europäischen Raum noch nicht ausreichend erforscht.

[1] Wenn in der vorliegenden Arbeit von Bewerbern, Studienanfängern, Probanden etc. gesprochen wird sind stets beide Geschlechter gleichermaßen gemeint.

Amerikanische Studien liefern Hinweise, dass Studierfähigkeitstests Frauen systematisch benachteiligen, indem sie deren Studienleistungen unterschätzen (Young & Kobrin, 2001). Diese Vermutung wird in der vorliegenden Arbeit aufgegriffen, mit dem Ziel herauszufinden, in welchem Ausmaß Studierfähigkeitstests den Studienerfolg von Frauen unterschätzen, inwiefern sich diese Befunde auch für ausgewählte deutschsprachige Studierfähigkeitstests zeigen und wie eine solche geschlechts-spezifische Unfairness in der Vorhersage erklärt werden kann.

2 Allgemeine Einführung

Das vorliegende Kapitel gibt einen umfassenden Überblick zum Thema Studierfähigkeitstests und deren Fairness. Zu Beginn wird aufgezeigt, was Studierfähigkeitstests genau messen und welche Testarten sich unterscheiden lassen. Anschließend werden die bedeutendsten amerikanischen Studierfähigkeitstests kurz vorgestellt und es wird der Einsatz von Studierfähigkeitstests in Deutschland beschrieben. Daraufhin wird erörtert, wie gut Studierfähigkeitstests den Studienerfolg vorhersagen, welche Merkmale einen fairen Test auszeichnen und inwiefern Studierfähigkeitstests diese Fairness-Anforderungen für Männer und Frauen erfüllen. Abschließend werden mögliche Ursachen für geschlechtsspezifische Unfairnessbefunde skizziert und die Forschungsfragen der vorliegenden Arbeit formuliert.

2.1 Messbereich und Testformen von Studierfähigkeitstests

Studierfähigkeitstests sollen diejenigen kognitiven Fähigkeiten messen, welche für ein erfolgreiches Studium erforderlich sind (z. B. Arnhold & Hachmeister, 2004; Rindermann & Oubaid, 1999). Dies geschieht gewöhnlich anhand von verbalen und numerischen *Reasoning-Aufgaben*, wie sie auch in Intelligenztests enthalten sind (Trost, 2003). Dementsprechend korrelieren allgemeine Verfahren zur Bestimmung der Intelligenz zu .5 bis .8 mit Studierfähigkeitstests (Frey & Detterman, 2004; Koenig, Frey & Detterman, 2008). Trotz dieser großen Überlappung der Messbereiche sieht Trost (2003, S. 14) wesentliche Unterschiede zwischen beiden Verfahren: Studierfähigkeits-tests differenzieren besser im oberen Leistungsbereich und fordern stärker den Umgang mit komplexen und umfangreichen Informationen, wohingegen Intelligenztests zusätzlich räumliches Vorstellungsvermögen erfassen.

Studierfähigkeitstests lassen sich gemeinhin in allgemeine und fachspezifische Verfahren unterscheiden, wobei letztere auf die spezifischen Anforderungen der

jeweiligen Studienfächer oder Studienfelder zugeschnitten sind (Deidesheimer Kreis, 1997, S. 90; Rindermann & Oubaid, 1999). Ein prominentes Beispiel für einen fachspezifischen Studierfähigkeitstest ist der deutsche Test für medizinische Studiengänge (TMS).

Während in Deutschland Studierfähigkeitstests strikt von Wissenstests abgegrenzt werden, ist der Übergang in anderen Ländern fließender (Haase, 2008). Vor allem amerikanische Studierfähigkeitstests enthalten auch Aufgaben, die studienrelevantes Wissen abfragen oder es stehen neben den *Reasoning-Testteilen* ergänzende Wissenstests, sogenannte *Subject-Tests,* zur Verfügung (Trost, 2003).

Im Fokus dieser Arbeit stehen sowohl allgemeine als auch fachspezifische Studierfähigkeitstests, welche vornehmlich als kognitive Fähigkeitstests konzipiert sind und sich klar von Wissenstests bzw. ergänzenden *Subject-Tests* abgrenzen lassen.

2.2 Die bedeutendsten amerikanischen Studierfähigkeitstests

Studierfähigkeitstests werden in Australien, Japan, Israel, Großbritannien und Schweden, sowie in vielen weiteren Ländern weltweit eingesetzt. In den USA haben Studierfähigkeitstests jedoch die längste Tradition. Bereits in den zwanziger Jahren setzten Universitäten an der amerikanischen Ostküste Tests zur Auswahl ihrer Studierenden ein (Trost, 2003; Zwick, 2002). Heutzutage spielen Studierfähigkeitstests an fast allen amerikanischen Hochschulen sowohl für die Zulassung zu *Undergraduate*- als auch zu *Graduate-Studiengängen* eine wichtige Rolle (Talento-Miller, 2008; Berry & Sackett, 2009).

Die Zulassung zum *Undergradute-Studium* wird von den beiden allgemeinen Studierfähigkeitstests SAT[2] und *American College Test (ACT)* dominiert, die in

[2] Früher stand SAT für Scholastic Aptitude Test und später für Scholastic Assessment Test, heute ist SAT ein eigenständiger Name.

Kombination mit der Abiturnote über die Vergabe der Studienplätze entscheiden. Allein den SAT absolvieren mehr als 1.5 Millionen Studienbewerber jährlich (College Board, 2011). Der SAT untergliedert sich in einen *Reasoning-Test* und einen ergänzenden *Subject-Test*, wohingegen der ACT stärker schulstoffspezifisches Wissen in Verbindung mit schlussfolgerndem Denken erfasst (Trost, 2003).

Für amerikanische *Graduate-Studiengänge* werden die Studierenden sowohl mit allgemeinen als auch mit fachspezifischen Studierfähigkeitstests ausgewählt. Am weitesten verbreitet ist die allgemeine *Graduate Record Examination (GRE)* mit etwa 500,000 Teilnehmern jährlich (Educational Testing Service, 2011). Ähnlich wie der SAT untergliedert sich dieser Test in einen allgemeinen Teil, der schlussfolgerndes Denken erfasst und in einen fachspezifischen *Subject-Test*.

Die bedeutendsten fachspezifischen *Graduate-Tests* in den USA sind der *Graduate Management Admission Test (GMAT)*, der *Law School Admission Test (LSAT)* und der *Medical College Admission Test (MCAT)*. All diese Tests beinhalten überwiegend kognitive Testaufgaben, nur der MCAT umfasst zusätzlich Wissenstestkomponenten. Eine detailliertere Beschreibung der einzelnen Verfahren ist beispielsweise bei Zwick (2002) oder Trost (2003) zu finden.

Die dargestellten Testverfahren haben sich seit Jahren in den USA etabliert und werden zunehmend auch an europäischen Hochschulen eingesetzt (z. B. an der Bocconi in Mailand oder dem Instituto de Empresa in Madrid). In Deutschland gewinnt insbesondere der GMAT für die Zulassung zu englischsprachigen Masterprogrammen an Bedeutung (Graduate Management Admission Council, 2012).

2.3 Studierfähigkeitstests in Deutschland

Im Gegensatz zu den USA wurde in Deutschland lange auf den Einsatz von Studierfähigkeitstest verzichtet. Erst mit dem Wunsch einer stärkeren Profilierung der

Hochschulen ist das Interesse an Studierfähigkeitstests gestiegen (Schmidt-Atzert & Krumm, 2006).

Mit dem siebten Gesetz zur Änderung des Hochschulrahmengesetzes im Jahr 2004 wurden rechtliche Rahmenbedingungen geschaffen, die den Hochschulen mehr Autonomie in der Vergabe ihrer Studienplätze einräumen. Von den zur Verfügung stehenden Studienplätzen in *bundesweit* zulassungsbeschränkten Fächern werden seitdem 60% durch die Hochschulen selbst vergeben (die restlichen 40% werden an Härtefälle, die Abiturbesten und nach der Dauer der Wartezeit verteilt; HRG, § 32 Abs. 2 und 3). Die Hochschulen können dabei neben schulischen Abschlussnoten auf Berufserfahrung, Interviews und *fachspezifische* Studierfähigkeitstests zurückgreifen (HRG, § 32 Abs. 3). *Örtlich* zulassungsbeschränkte Studienplätze werden über hochschuleigene Auswahlverfahren vergeben, deren Rahmenbedingungen die jeweiligen Gesetzgebungen auf Landesebene klären, welche jedoch meist an die bundesweiten Regelungen angelehnt sind. Forderungen nach einem allgemeinen, fächerübergreifenden Studierfähigkeitstest, ähnlich dem SAT (z. B. Rindermann, 2005), sind in der deutschen Gesetzeslage derzeit unbeachtet.

Der wohl bekannteste deutsche Studierfähigkeitstest ist der TMS, welcher seit 2008 wieder bundeslandübergreifend für die Zulassung in den Studiengängen Humanmedizin, Zahnmedizin und Veterinärmedizin verwendet wird, nachdem er bereits in den Jahren 1986 bis 1996 eingesetzt wurde. Aktuell berücksichtigen 14 medizinische und sieben zahnmedizinische Fakultäten den Test in ihrem Auswahlverfahren (Stand Wintersemester 2011/2012). Der TMS misst fachspezifische kognitive Fähigkeiten, wie schlussfolgerndes Denken, visuelle Informationsverarbeitung und Merkfähigkeit (Maichle & Meyer, 1997). Die Teilnahme am TMS ist nicht obligatorisch, sie kann jedoch die Chance auf einen Studienplatz im Rahmen der hochschuleigenen Auswahlverfahren erhöhen.

Abgesehen von den medizinischen Fakultäten setzen derzeit vor allem private Hochschulen sowie Fachhochschulen für die Zulassung zu wirtschaftswissenschaftlichen

Studiengängen Studierfähigkeitstests ein (Zimmerhofer & Trost, 2008). Die meisten dieser Tests werden von der ITB Consulting GmbH entwickelt, welche auch den TMS betreut. Analog zum TMS sind die wirtschaftswissenschaftlichen Tests fachspezifische kognitive Fähigkeitstests. Weitere lokale Testverfahren, wie beispielsweise für das Fach Psychologie an der Universitäten Berlin (Formazin, Schroeders, Köller, Wilhelm & Westmeyer, 2011), wurden in jüngster Zeit erprobt, ihre Verwendung hat sich jedoch (noch) nicht etabliert.

Neue Einsatzmöglichkeiten für Studierfähigkeitstests in Deutschland ergeben sich aus der Umstellung auf Bachelor- und Masterabschlüsse im Rahmen des Bologna-Prozesses. Die entstehenden interdisziplinären Masterprogramme bringen neue Herausforderungen in der Studierendenauswahl mit sich, da die Abschlussnoten aus unterschiedlichen Bachelor-Studiengängen nicht zwingend vergleichbar sind (Zimmerhofer & Trost, 2008). Studierfähigkeitstests bieten hingegen die Möglichkeit, das kognitive Potential der Bewerber objektiv zu bestimmen.

Neben der Auswahl von Studierenden werden Studierfähigkeitstests in Deutschland zunehmend im Rahmen der Studienberatung eingesetzt. Die Bundesagentur für Arbeit bietet *Studienfeldbezogene Beratungstests* an (Psychologischer Dienst, 2011) und immer mehr Hochschulen bauen auf Self-Assessment-Angebote, die auch Studierfähigkeitstest-Komponenten enthalten, wie beispielsweise die RWTH Aachen (Zimmerhofer, Heukamp & Hornke, 2006) und die Universität Frankfurt (Kubinger, Moosbrugger, Frebort, Jonkisz & Reiß, 2007). Ziel dieser Verfahren ist es, Studieninteressierten eine Rückmeldung darüber zu geben, in welchem Ausmaß ihr angestrebtes Studienfeld ihren Fähigkeiten entspricht (Kubinger, Frebort & Müller, 2012; Psychologischer Dienst, 2011). Die rückgemeldeten Informationen sollen bereits vor dem eigentlichen Auswahlprozess der Hochschulen eine Selbstselektion der Bewerber ermöglichen und den tatsächlichen Auswahlaufwand deutlich reduzieren (Amelang & Funke, 2005; Jäger, 2005).

2.4 Die prädiktive Validität von Studierfähigkeitstests

Im Vergleich zu anderen diagnostischen Verfahren wird bei Studierfähigkeitstests einer guten Kriteriumsvalidierung oft mehr Bedeutung beigemessen als einer genauen Konstruktvalidierung. Der Studienerfolg gilt als Kriterium der Wahl, ist jedoch gleichzeitig ein mehrschichtiges und facettenreiches Konstrukt (Konegen-Grenier, 2001). Rindermann und Oubaid (1999) schlagen eine Aufschlüsselung in folgende sechs Kriterien vor: Studienabschluss, Studiennoten, Studiendauer, Studienzufriedenheit, berufsqualifizierende Kompetenzen und Berufserfolg. Im Rahmen von Validierungsstudien werden selten alle genannten Facetten berücksichtigt. Meist wird auf die Noten aus dem Grundstudium als Erfolgskriterium zurückgegriffen (Rindermann & Oubaid, 1999; Wissenschaftsrat, 2004), wie auch bei den folgenden Untersuchungen.

Zahlreiche Metaanalysen und groß angelegte Evaluationsstudien belegen Studierfähigkeitstests eine zufriedenstellende Vorhersagegüte. Im deutschsprachigen Raum beträgt die mittlere Validität fachspezifischer Studierfähigkeitstest .31 und nach Korrektur für die Reliabilität der Studiennoten und die Variabilitätseinschränkung der Testergebnisse sogar .48 (Hell et al., 2007). In den USA erreicht der SAT eine korrigierte Validität im Bereich von .54 (Patterson, Mattern & Kobrin, 2009; Patterson & Mattern, 2011) und *Graduate-Tests* erzielen Korrelationen zwischen .41 und .59 (Kuncel & Hezlett, 2007).

Zur besseren Einordnung der beschriebenen Validitäten ist es hilfreich diese mit Kennwerten anderer Hochschulzulassungsverfahren zu vergleichen. Auswahlgespräche zeigen beispielsweise mit einer korrigierten Validität von .16 eine deutlich geringere Vorhersagegüte, wobei strukturierte Auswahlgespräche besser abschneiden als unstrukturierte (Hell, Trapmann, Weigand & Schuler, 2007). Schulnoten gelten als bester Einzelprädiktor zur Vorhersage von Studienerfolg (Hell, Trapmann & Schuler, 2008). Mit mittleren korrigierten Validitätskoeffizienten von .53 bis .54 ist die prädiktive Validität der Abiturnote bzw. der *High-School-Note* meist etwas höher als die der Studierfähigkeitstests (Trapmann, Hell, Weigand & Schuler, 2007; Kobrin, Patterson,

Shaw, Mattern & Barbuti, 2008; Patterson et al., 2009; Patterson & Mattern, 2011). Die Hinzunahme von Studierfähigkeitstests zur Abiturnote verbessert die Validität jedoch nochmals bedeutsam (Hell et al., 2008; Kobrin et al., 2008).

Die beschriebenen Studien machen deutlich, dass Studierfähigkeitstests alleine und in Verbindung mit der Abiturnote eine hohe prädiktive Validität aufweisen. Offen bleibt jedoch die Frage, ob Studierfähigkeitstests gleichzeitig allen Teilnehmergruppen gegenüber fair sind.

2.5 Welche Kriterien sollte ein fairer Studierfähigkeitstest erfüllen?

In der Vergangenheit wurde kontrovers diskutiert, wie die Fairness eines Tests zu operationalisieren ist (Meade & Fetzer, 2009). Hierbei wurden sowohl statistische, psychometrische als auch soziale Aspekte betrachtet (Society for Industrial and Organizational Psychology [SIOP], 2003). Die *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME], 1999, S. 74-80) fassen die Debatte zusammen, indem sie eine Unterscheidung in folgende vier Fairness-Aspekte vorschlagen:

1. Gleichheit im Test-Outcome, d. h. gleiche Zulassungsraten für alle Teilnehmergruppen.

2. Gleiche Behandlung aller Testteilnehmer, d. h. alle Testteilnehmer sollten die gleichen Testbedingungen und Feedbackinformationen erhalten.

3. Gleiche Lerngelegenheiten, d. h. alle Teilnehmergruppen sollten vorab die gleichen Möglichkeiten haben sich testrelevantes Wissen anzueignen.

4. Lack of bias. Dieser schwer zu übersetzende Terminus (am ehesten: *unverzerrte Vorhersage*) bezieht sich auf die statistischen Gütemerkmale der Tests und

impliziert die Abwesenheit von *differential item functioning* (DIF), differenziellen Validitäten und differenzieller Prädiktion.

Die *Standards* (AERA, APA & NCME, 1999) betonen die Aspekte zwei bis vier und stehen dem ersten Fairness-Aspekt sehr kritisch gegenüber. Sie weisen darauf hin, dass Mittelwertsunterschiede im Test nicht unbedingt durch Quotenregelungen ausgeglichen werden sollten, da Mittelwertsdifferenzen nicht zwingend einen unfairen Test implizieren, sondern vielfältige Ursachen haben können.

Weiterhin kann diskutiert werden, für welche Teilnehmergruppen die beschriebenen Fairness-Kriterien erfüllt sein sollten. In den USA stehen ethnische Minderheiten im Fokus (z. B. Bridgeman, McCamley-Jenkins & Ervin, 2000; Zwick & Himelfarb, 2011), aber auch Unterschiede zwischen Männern und Frauen werden umfassend thematisiert (z. B. Bridgeman et al., 2000; Young & Kobrin, 2001). In Deutschland und in Österreich hat die Genderfairness in den letzten Jahren ebenfalls an Interesse gewonnen, da sich die Befunde häufen, dass Frauen im Mittel schlechtere Testergebnisse erzielen als Männer (Blum, 1997; Dlugosch, 2005; Spiel, Schober & Litzenberger, 2008). Vorliegende Arbeit knüpft an diese Befunde an und beschäftigt sich mit der Genderfairness von Studierfähigkeitstests und den sich aus den Ergebnissen ergebenen Implikationen für die Beratung und Auswahl von Studierenden.

2.6 Inwiefern sind Studierfähigkeitstests genderfair?

Im Folgenden wird erörtert, in welchem Umfang Studierfähigkeitstests die von den *Standards* (AERA, APA & NCME, 1999) geforderten Fairness-Aspekte (vgl. Abschnitt 2.5) für Männer und Frauen erfüllen.

Gleiche Behandlung aller Testteilnehmer

In den letzten Jahren wurde intensiv darauf geachtet, dass allen Teilnehmern von Studierfähigkeitstests kostenfreie Übungsmaterialien zur Verfügung stehen, so dass kommerziellen Anbietern wenig Raum für kostenpflichtige Vorbereitungskurse gegeben wird. Beispielsweise können für den TMS kostenlos Übungsaufgaben im Internet heruntergeladen werden (ITB Consulting GmbH, 2012). Benachteiligungen im Sinne einer Ungleichbehandlung bezüglich der Geschlechter sind nicht bekannt.

Gleiche Lerngelegenheiten

In den USA existieren deutliche qualitative Unterschiede zwischen den *High-Schools* (Berkowitz & Hoekstra, 2011; Pike & Saupe, 2002). Immigranten besuchen häufiger qualitativ schlechte Schulen, wodurch sie geringere Chancen auf Bildung haben und somit bei Wissenstests benachteiligt werden (Fletcher & Tienda, 2010). Dies betrifft jedoch beide Geschlechter gleichermaßen, so dass von keiner geschlechts-spezifischen Unfairness gesprochen werden kann.

Lack of bias

Kein DIF. Geschlechtsspezifisches DIF liegt vor, wenn die Wahrscheinlichkeit der korrekten Lösung eines Items für Männer und Frauen unterschiedlich groß ist, obwohl beide Geschlechter die gleiche Ausprägung des zugrundeliegenden Konstrukts besitzen. Für den SAT existieren ausführliche DIF Untersuchungen (z. B. Curley & Schmitt, 1993; Schmitt & Dorans, 1990). Potentielle Items für neuere Testversionen werden bereits während der Testkonstruktion überprüft, so dass unfaire Items erst gar nicht in den eigentlichen Test gelangen (Mattern, Patterson, Shaw, Kobrin & Barbuti, 2008). Andere Testanbieter führen jedoch keine so umfangreichen Vorerprobungen ihrer Items durch, so dass geschlechtsspezifisches DIF nicht ausgeschlossen werden kann.

Keine differenziellen Validitäten. Ob Studierfähigkeitstests die Studienleistungen von Männern und Frauen gleich gut vorhersagen, ist umfassend erforscht. Der österreichische Eignungstest für das Medizinstudium korreliert für Männer zu .50 und für Frauen zu .53 mit dem Kriterium Studienerfolg (Werte sind korrigiert für Varianzeinschränkung; Hänsgen et al., 2008). Der SAT zeigt ebenfalls geringfügig höhere Validitäten für Frauen als für Männer (Ramist, Lewis & McCamley-Jenkins, 1994; Young & Kobrin, 2001). Insgesamt sind die geschlechtsspezifischen Validitätsunterschiede über verschiedene Tests hinweg robust, jedoch gleichzeitig von sehr geringem Ausmaß.

Keine differenzielle Prognose. Eine differenzielle Prognose liegt vor, wenn durch die Ergebnisse im Studierfähigkeitstest die Studienleistungen eines Geschlechts systematisch über- oder unterschätzt werden. Dies kann zur Folge haben, dass bestimme Bewerber nicht zum Studium zugelassen werden, obwohl sie erfolgreich die geforderten Studienleistungen erbringen würden (Huff, Koenig, Treptau & Sireci, 1999). Forschungsarbeiten aus den USA deuten darauf hin, dass Studierfähigkeitstests den Studienerfolg von Frauen unterschätzen (Young & Kobrin, 2001), es ist jedoch unklar, inwiefern diese Befunde generalisierbar sind. Im deutschsprachigen Raum gibt es zur geschlechtsspezifischen differenziellen Prognose von Studierfähigkeitstests nur sehr wenige Untersuchungen. Für den TMS zeigt sich im Studienfach Medizin eine geringe Unterschätzung der Leistungen von Frauen (Nauels & Meyer, 1997) und auch der Studierfähigkeitstest der privaten Bucerius Law School im Studienfach Jura unterschätzt tendenziell den Studienerfolg der Frauen (Dlugosch, 2005). Insgesamt ist der Forschungsstand in Deutschland in Bezug auf die geschlechtsspezifische differenzielle Prognose von Studierfähigkeitstests jedoch defizitär. Auch ist unklar, ob alle Leistungsbereiche gleich stark von der Über- bzw. Unterschätzung betroffen sind. Eine detaillierte Beschreibung, wie differenzielle Prognose gemessen werden kann, folgt in Abschnitt 3.2.

Zusammenfassend lässt sich festhalten, dass Studierfähigkeitstests viele der von den *Standards* (AERA, APA & NCME, 1999) geforderten Fairness-Kriterien für Männer und Frauen erfüllen, jedoch im Bezug auf DIF und die differenzielle Prognose noch erheblicher Forschungsbedarf besteht. Die vorliegende Arbeit greift die Frage nach der geschlechtsspezifischen differenziellen Prognose von Studierfähigkeitstests auf und betrachtet genauer, aus welchen Gründen sich möglicherweise eine Unterschätzung der Studienleistungen von Frauen zeigt.

2.7 Mögliche Gründe für eine Unterschätzung der Studienleistungen von Frauen durch Studierfähigkeitstests

Ursachen der geschlechtsspezifischen differenziellen Prognose von Studierfähigkeitstests wurden in der Vergangenheit kontrovers diskutiert. Uneinheitliche Forschungsergebnisse lassen vermuten, dass nicht alle Studierfähigkeitstests gleich stark betroffen sind, sondern das Testalter, die Testart oder die Testinhalte möglicherweise entscheidend sind für das Ausmaß der differenziellen Prognose.

Alternativ wird argumentiert, dass Frauen ein anderes Kurswahlverhalten zeigen als Männer und damit die Studiennoten von Frauen und Männern nicht vergleichbar sind (Berry & Sackett, 2009). Frauen wählen einfachere Fächer und erhalten somit bessere Noten als ihre männlichen Kommilitonen. Fasst man diese eigentlich nicht vergleichbaren Noten zusammen, kann man zu der (falschen) Schlussfolgerung kommen, dass Studierfähigkeitstests die Studiennoten von Frauen unterschätzen (Elliott & Strenta, 1988; Hewitt & Goldman, 1975).

Geschlechtsspezifische Persönlichkeitsunterschiede und damit zusammenhängendes Studienverhalten werden ebenfalls für die Unterschätzung des Studienerfolgs der Frauen verantwortlich gemacht (Sackett, Borneman & Connelly, 2008). Studentinnen sind demnach motivierter und zeigen mehr Selbstdisziplin, wodurch sie bei

gleichen kognitiven Kapazitäten bessere Noten erzielen als ihre männlichen Kommilitonen (Zwick, 2002, S. 151). Die Abiturnote scheint Persönlichkeitsmerkmale zu berücksichtigen und deshalb den Studienerfolg von Frauen weniger zu unterschätzen (Mattern et al., 2008). Welche Persönlichkeitsmerkmale jedoch entscheidend sind für den Studienerfolg von Frauen und durch die Abiturnote miterfasst werden, ist unklar.

Ebenso wird angenommen, dass der Einfluss von *stereotype threat* während der Testsituation Frauen schlechter abschneiden lässt als sie es eigentlich könnten (Spencer, Steele & Quinn, 1999; Steele, 1997). Bis heute ist nicht eindeutig geklärt, welche der genannten Faktoren entscheidend sind für eine mögliche Unterschätzung der Studienleistungen von Frauen durch Studierfähigkeitstests.

2.8 Ziele der Dissertation

Studierfähigkeitstests haben einen großen Einfluss auf den Studien- und Berufsweg von Studieninteressierten. Trotz dieser großen Lenkungswirkung ist nicht hinreichend geklärt, ob Studierfähigkeitstests die Studienleistungen von Frauen systematisch unterschätzen und ob nur bestimmte Tests oder nur bestimmte Leistungsbereiche davon betroffen sind. Im Vergleich zu den USA ist die Forschungslage in Deutschland hierzu besonders defizitär. Darüber hinaus sind die Gründe für mögliche Unfairnessbefunde unklar. Verschiedene Erklärungsansätze werden in der Literatur diskutiert, aber keine dieser Annahmen ist hinreichend empirisch belegt.

Da für die Studierendenauswahl gewöhnlich Studierfähigkeitstests in Kombination mit der Abiturleistung als Zulassungskriterium herangezogen werden, ist für die Praxis zusätzlich von Interesse, wie fair die Vorhersage für die Kombination aus Test und Abiturnote ausfällt.

Anknüpfend an die beschriebene unbefriedigende Forschungslage zur geschlechtsspezifischen Über- und Unterschätzung des Studienerfolgs durch Studierfähigkeitstests ist das Ziel dieser Arbeit, neue Erkenntnisse zu folgenden Fragen zu gewinnen:

I. Besteht eine generalisierbare Unterschätzung (im Sinne der differenziellen Prognose) der Studienleistungen von Frauen durch Studierfähigkeitstests?

II. Wirkt sich die Berücksichtigung der Abiturnote günstig auf die Vorhersage-Fairness aus – und wenn ja, warum?

III. Zeigt sich eine Unterschätzung der Studienleistungen von Frauen auch für ausgewählte deutschsprachige Studierfähigkeitstests?

IV. Was sind die Ursachen für mögliche Unfairnessbefunde? Lassen sich die bestehenden Erklärungsansätze (vgl. Abschnitt 2.7) belegen und/oder durch neue Befunde erweitern?

Die Beantwortung dieser Fragen erfolgt anhand folgender drei Studien:

Die erste Studie fasst den aktuellen, internationalen Forschungsstand zur geschlechtsspezifischen differenziellen Prognose von Studienleistungen durch Studierfähigkeitstests zusammen. Hierfür werden metaanalytische Techniken verwendet, um Aussagen über die Generalisierbarkeit bislang gewonnener Einzelbefunde abzuleiten. Darüber hinaus werden durch Moderatorenanalysen Bedingungen identifiziert, unter denen Studierfähigkeitstests zu einer unfairen Prognose gelangen. Zusätzlich wird betrachtet, ob sich die geschlechtsspezifische differenzielle Prognose verändert, wenn das Vorhersagemodell durch die Abiturnote ergänzt wird (Kapitel 3).

Aufbauend auf den Ergebnissen der ersten Studie untersucht die zweite Studie, welche Persönlichkeitsmerkmale neben den kognitiven Fähigkeiten den Abiturerfolg von Frauen und Männern determinieren und damit das bessere Abschneiden von Frauen

in der Schule erklären. Diese Ergebnisse liefern Hinweise, warum die Abiturnote möglicherweise geschlechtsspezifische Studienleistungen fairer vorhersagt als Studierfähigkeitstests (Kapitel 4).

In der dritten Studie wird in einem längsschnittlichen Studiendesign die geschlechtsspezifische differenzielle Prognose von zwei deutschsprachigen, fachspezifischen Studierfähigkeitstests untersucht. Darüber hinaus wird analysiert, ob die differenzielle Prognose alle Leistungsbereiche gleich stark betrifft oder ob eine Unterschätzung der Studienleistungen von Frauen nur bei sehr strengen Selektionsquoten auftritt. Zusätzlich werden neue Erkenntnisse über die Ursache für auftretende Unfairnessbefunde gewonnen, indem der Einfluss von Persönlichkeitsmerkmalen überprüft wird (Kapitel 5).

3 Sex-Specific Differential Prediction of College Admission Tests: A Meta-Analysis

[Geschlechtsspezifische differenzielle Prädiktion von Studierfähigkeitstests: Eine Metaanalyse]

Franziska Fischer,

Johannes Schult &

Benedikt Hell

Universität Konstanz

Fischer, F., Schult, J., & Hell, B. (2013). Sex-specific differential prediction of college admission tests: A meta-analysis. *Journal of Educational Psychology*, *105*(2), 478-488. doi: 10.1037/a0031956

3.1 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 as well as criterion differences 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.

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

3.2 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 validity for success at college is revealed by several meta-analyses: The raw correlation between the SAT and first-year college or university grade point averages (GPA) is .35. The correlation increases to .53 after correction for range restriction[3] (Kobrin et al., 2008). Similar results have been found for the GMAT (Kuncel, Credé, & Thomas, 2007), the GRE (Kuncel, Hezlett, & Ones, 2001; Kuncel, Wee, Serafin, & Hezlett, 2010), and subject specific admission tests in German speaking countries (Hell et al., 2007). Although predictive validity is necessary for high stakes testing, it is not sufficient for its fairness (AERA, APA, & NCME, 1999).

Another aspect which must not be neglected is the discrimination against subgroups (e.g., Meade & Fetzer, 2009). 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. As a possible consequence females with the same academic potential as their male classmates are less likely to be admitted to a college or university if admission tests are the only criterion for admission.

Efforts to summarize the literature in this field have are more than ten years old and restricted with regard to content and method. The present study overcomes these

[3] 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 (Thorndike, 1949). 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).

limitations by providing an up-to-date meta-analysis. International research results respecting 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.

Test Fairness and Test Bias in Predicting Subgroups

Test fairness and *test bias* have been studied intensively in the past. Regarding the definition of these two issues there has been some disagreement. 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 (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) used 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*.

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 et al., 1999).

How to Measure Differential Prediction

Analyzing differences in regression equations. The first method for testing the differential prediction hypothesis was introduced by Gulliksen and Wilks (1950). They 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; (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., Bridgeman & Wendler, 1991; Cleary, 1968).

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, Bobko, Mosier, & Hannan, 1978). The corresponding formula is

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1 X_2, \qquad (1)$$

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. To arrive at a conclusion about over- and underprediction is more complicated in the presence of different slopes. Regression lines may cross, and therefore an investigation of different test score sections is desirable.

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; Clark & Grandy, 1984; Talento-Miller, 2008).

In large-scale studies, group-specific residuals are preferred to the significance tests of the Cleary approach. The reason for this being that statistical tests are prone to indicate significance when large samples are analyzed due to the large test power (e.g., Cohen, 1988). The College Board, for example, regularly publishes SAT studies reporting residuals for more than 100,000 students (e.g., Bridgeman, Pollack, & Burton, 2008; Mattern et al., 2008; Patterson, Mattern, & Kobrin, 2009).

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.

The first study is an unpublished meta-analysis that summarizes differential prediction effects associated with standardized achievement tests (Sanber & Millman, 1987). The authors aggregated 38 studies including 147 samples. Results show 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, however, 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.

The second study is an extensive summary by Young and Kobrin in 2001. They review the literature on differential prediction in American college admission. The summary 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 are included. For example, residuals from male regression lines as well as structural equation modelling parameters are summarized and studies using a combination of test scores and high school grades as predictor are not considered separately.

In summary, one can say that no independent up-to-date meta-analysis about the differential prediction of college and graduate admission tests exists.

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 suggests that combining grades and test scores yields less bias than test scores alone (e.g., Mattern et al., 2008).

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 databases that list primary studies from all over the

world. We also include recent findings, since many large-scale studies have been published in the last decade. Last, but not least, we apply metaanalytic 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, we focus on the following:

First, we look at publication and sample properties. Changes over time are examined to investigate whether differential prediction is just a problem of older tests. 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) and it is investigated if differential prediction is related to cultural differences between samples.

Second, 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 et al., 2008; Patterson et al., 2009). We also test if differential prediction is related to test score differences between men and women, differences in college grades, and average time span between predictor and criterion assessment. These analyses allow conclusions whether over- and underprediction is more closely associated with a bias in the test or a bias in the criterion (Meade & Fetzer, 2009; Meade & Tonidandel, 2010).

Finally, we look at course taking patterns. It was argued that females tend to enroll in less stringent courses with more lenient grading systems (Alon & Gelbgiser, 2011; Conger & Long, 2010). Correcting for differences in grading standards or course taking patterns reduced underprediction of women (Bridgeman et al., 2000; Ramist et al., 1994; Willingham, Pollack, & Lewis, 2002). But there also exists evidence that

underprediction of women's grades persists after controlling for gender differences in fields of study and for sample selection bias (Leonard & Jiang, 1999).

3.3 Method

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.

Inclusion Criteria

Each of the potential articles was evaluated for inclusion based on the following criteria. First, the study had to examine the prediction of men's and women's college performance by an admission test. Alternatively the prediction had to be based on a combination of admission test results and previous grades. Second, 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. Third, 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. This procedure helps to avoid a duplicate study effect (Wood, 2008).

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 like author, sample size, and name of the admission test are presented in Tables 3.1 and 3.2. Given that the studies often contained several independent samples, the distinguishing characteristics of each sample appear in the second column of the tables.

Coding of Study Variables

Data of independent samples such as different colleges 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. First, the whole test was used as predictor instead of test parts. Second, the criterion with the biggest sample size was chosen. In ambiguous cases, the first year GPA was analyzed instead of later earned grades.

Following aspects were coded as potential moderators: publication year, country of study origin, country of sample origin, 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.

The first and the second authors coded all of the studies independently. Both were familiar with the field of study and had created the coding scheme. The initial interrater 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.

Table 3.1

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

					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 Psychodiagnos-tics 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

					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 and 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 3.2

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

					Significant intercept 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	-	yes
	College 2a	SAT-M	Calculus grades	106	-	yes
	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	-	yes
Calkins & Whitworth, 1974		SAT	GPA attained for the first 30 or less college hours	3,237	-	yes
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	yes	-
	2001 sample	Test of the BLS	GPA after 2 years	91	no	-
Hewitt & Goldman, 1975	Los Angeles campus	SAT	GPA	-	yes	-
	Davis campus	SAT	GPA	-	yes	-
	Irvine campus	SAT	GPA	-	yes	-
	San Diego campus	SAT	GPA	-	yes	-
Hogrebe et al., 1983		SAT	FGPA	345	-	yes
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	yes	yes
	Veterinary medicine students	TMS	Examination after 1 year	2,391	yes	yes
	Dentistry students	TMS	Examination after 1 year	5,221	yes	yes
Patton, 1998	Biology students	ACT	Cumulative college GPA	195	no	-
	English students	ACT	Cumulative college GPA	254	yes	-
	Finance students	ACT	Cumulative college GPA	257	no	-

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

Note. Dash indicates that the data is not provided for the sample and could not be calculated. 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.

Analytical Procedures

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

$$t = [(\overline{E}_{men} - \overline{E}_{women}) / SD] \cdot \sqrt{N} \ . \qquad (2)$$

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

$$N_{men} \cdot \overline{E}_{men} + N_{women} \cdot \overline{E}_{women} = 0 \ . \qquad (3)$$

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:

$$t_j = [(\overline{E}_j - 0) / SD] \cdot \sqrt{N_j} \qquad (j = men, women) \ .^{4} \qquad (4)$$

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

$$d_j = (\overline{E}_j - 0 \ ) / SD \qquad (j = men, women) \ . \qquad (5)$$

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

$$SD = \sqrt{S_y^2 \cdot (1 - R_{xy}^2)} \ , \qquad (6)$$

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

[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.

After calculating effect sizes within each sample, we separately accumulated the *d*-values for men and women. For each gender a bare-bones meta-analysis was 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). The corresponding formula for the weighted average effect size was

$$\bar{d}_j = \sum w_{ji} d_{ji} / \sum w_{ji} \quad (j = men, women; i = sample) \qquad (7)$$

with w_i as the weight for the *i*th 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 in the studies with regard to the mean residuals.

We calculated 95% confidence intervals, and 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 (1994) based on a random-effects model. Due to the fact that the *Q*-test often does not have an acceptable power, heterogeneity tends to remain undetected (Schulze, 2004). We therefore also performed the heterogeneity test 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 of conducting a meta-analysis of regression equations (for a detailed discussion see Aguinis, Culpepper, & Pierce, 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 in primary studies.

A new approach recommended using the semipartial 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 1). The interaction term (sex*test score) represents different gradients for men and women. 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). Again, our identified studies rarely reported this information (especially the correlation). 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-Roman, 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.

3.4 Results

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.[5]

The funnel plots of the four data sets resemble a relatively symmetric inverted funnel, indicating the absence of publication bias (Light & Pillemer, 1984; see Figure 3.1). Also 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 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. As shown in Figure 3.1, there are two outliers in the funnel identified as asymmetric. Both samples show a very low percentage of men (less than 28%). Following equation 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 3.3 shows the corresponding mean effect sizes for women ($d_{female} = .14$, indicating underprediction) and men ($d_{male} = -.16$, indicating overprediction). Due to the fact that neither confidence nor credibility intervals overlap zero, the results can be generalized (Schmidt & Hunter, 1982). Heterogeneity analysis based on a random-effects model reveals homogenous effect sizes for women

[5] We also calculated mean effect sizes across the studies, without the four large-scale studies ($N > 5{,}000$). The results were substantially the same.

$Q_{random}(54) = 66.83$, $p = .113$ and for men $Q_{random}(54) = 57.15$, $p = .359$. According to a fixed effect model the effect sizes are heterogeneous for women, $Q_{fix}(54) = 111.31$, $p < .001$ and homogenous for men $Q_{fix}(54) = 69.49$, $p = .076$.

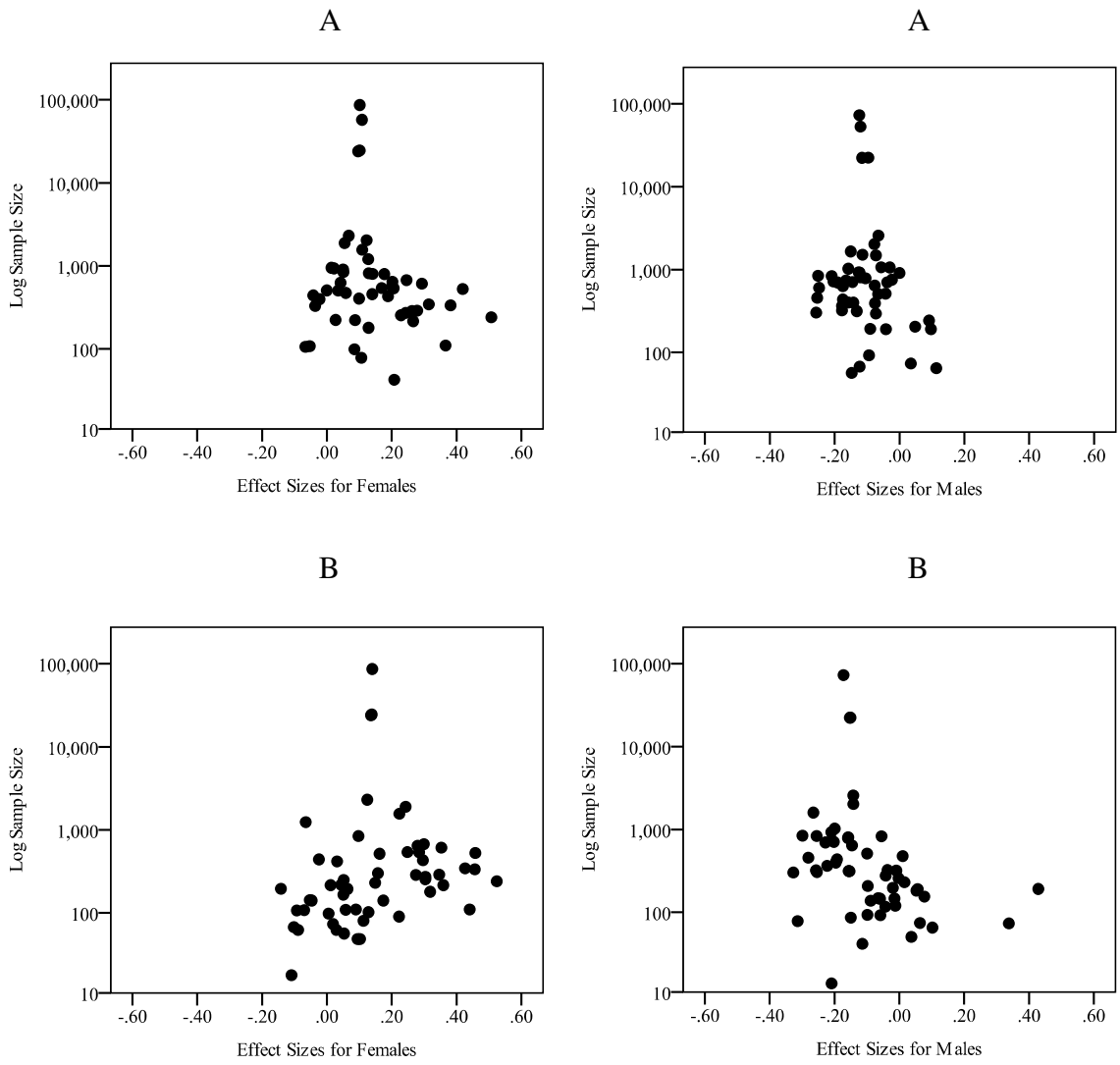

Figure 3.1. Funnel plots of female's and male's effect sizes against logarithms of sample sizes. The effect sizes represent the magnitude of over- or underprediction. A: using admission test as predictor, B: using admission test and HGPA/UGPA as predictor

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_{female} = .11 and d_{male} = -.12; see Table 3.3). 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_{random} (50) = 73.31, p < .05 and Q_{fix}(50) = 76.75, p < .01 and homogeneous for men $Q_{random/fix}$(51) = 49.76, p = .523 – for random as well as for fixed effects models.

Table 3.3

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.

Moderator analysis. Effect sizes are heterogeneous for women based on admission tests as the sole predictor (according to a fixed effects model). To explain this effect size distribution for women we conducted analyses for potential moderators.

Test type is a significant moderator, $Q_{between}$(5) = 76.97, p < .001. The effect sizes for the different tests where more than one sample was available are displayed in Table 3.4. The underprediction of women's academic performance is negligible for the graduate admission tests, GRE and MCAT, and small to moderate for the undergraduate

admission tests, SAT[6] (d_{female} = .14) and ACT (d_{female} = .30). For the Miller Analogies Test and the GMAT only one sample is available. The corresponding effect sizes can be found in Table 1.

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_{female} = .17, k = 4, N_{female} = 135,144, 95% CI [.15, .19], 90% CRI [.15, .19]) than the verbal section (d_{female} = .10, k = 4, N_{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 publication year and time span 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 3.5.

The variables country of study origin and country of sample origin do not have enough variation to test their influence. The included samples almost exclusively originate from the United States. For the other potential moderator variables (age and course choice), there was not enough data from the primary studies for analysis.

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

Table 3.4

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; d_f = effect size for females based on admission test as predictor, whereas positive effect sizes indicate underprediction; CI = confidence interval; CRI = credibility interval; ACT = American College Test; GRE = Graduate Record Examination; MCAT = Medical College Admission Test.

Table 3.5

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	14	.375	.14	.265
Criterion differences	32	-.174	.03	.492
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 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).

Differences in Group Regression Equations

As described in the section s*ummarizing 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_{mean} = 2,032) than the average sample size of the studies reporting no differences (N_{mean} = 573). This is not a surprise since significance depends, besides other factors, on sample size.

3.5 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, Beaty, Boik, and Pierce (2005) showed that this occurs frequently in studies of differential prediction.

When the effect sizes are transferred onto a four-point grading scale[7], 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_{female} = .11, d_{male} = -.12)[8]. 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. But, 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 prediction 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.

[7] Plugging in the mean standard deviation of residuals of the studies with the largest sample sizes (Patterson et al., 2009; Bridgeman et al., 2008).

[8] 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_{female} = .07, k = 24, N_{female} = 144,383, 95% CI [.06, .09], 90% CRI [.03, .12], $Q(23)$ = 50.99, p < .001) and very small overprediction for men (d_{male} = -.08, k = 24, N_{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.

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. In other words, publication year is not a significant moderator.

There are three particular reasons for over- and underprediction: bias in the test, bias in the criterion and omitted variables (e.g., Meade & Fetzer, 2009; Meade & Tonidandel, 2010). Our meta-analysis shows that the underprediction of women is associated neither with predictor differences (possibly indicating a bias in the test) nor with criterion differences (possibly indicating a bias in the criterion). Higher mean test scores of males are not related to females' underprediction. Consequently, disposing test score differences, by restructuring a test, does not necessarily reduce underprediction. Similarly, females' better grades at college (e.g., Ellis et al., 2008, p. 278) are not an indicator for their underprediction.

Subsequent moderator findings demonstrate that different admission tests show different levels of over- and underprediction. The underprediction of women's academic performance is small for the SAT and ACT, and it is negligible for the GRE and MCAT. It has to be mentioned that the results for the ACT and MCAT are limited by the small number of available studies; the results are not independent of the corresponding samples. 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 presented findings challenge the assumption that women's academic performance is underpredicted because they experience *stereotype threat* during the test execution. Stereotype threat 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). A meta-analysis of experimental studies showed that the stereotype threat decreases the test performance of women (Nguyen &

Ryan, 2008). Still, the examination of stereotype threat in real world settings is difficult and 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.

The underprediction linked to undergraduate tests might be explained more accurately 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 is reduced (Stricker, Rock, & Burton, 1993; Veldman, 1968). As graduate students are a more *selective sample*, they possibly show more homogenous personality characteristics and skills across the sexes than high school alumni.

We also found differences between test components. The SAT verbal section shows less underprediction than the mathematics section. Several explanations for these differences are discussed in the scientific circles. It has been argued that SAT math questions refer to situations more familiar to boys like sports teams and male recreational activities and therefore disadvantage women (Zwick, 2002). This hypothesis is challenged by Grand, Ryan, Schmitt, and Hmurovic (2011) who find that adding male stereotyped job context to test items does not negatively impact the performance of women. However, since test items are changed on a regular basis, it is very difficult to prove their influence.

Another explanation could be the *multiple-choice format.* This format 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 (Bolger & Kellaghan, 1990; Bridgeman & Lewis, 1994; Lindberg, Hyde, Petersen, & Linn, 2010; Murphy, 1982). Lack of time might be responsible for these differences (Goldstein, Haldane, & Mitchell, 1990). This assumption is supported by results which indicate that females omit more items, especially towards the end of the

test. Alternatively, the increasing difficulty of the test items can be responsible for this effect (Mäkitalo, 1996; Åberg-Bengtsson, 1999).

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

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.

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 included 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. At least some of the differential prediction results could be an artifact of sample size.

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.

Using the concept of one common regression line yields a total mean residual of zero. The mean residual for women is not independent from the male residual and their sample sizes respectively (see equation 3). As a consequence of this statistical fact, minorities reach more extreme mean residuals than the corresponding majority. Meta-analyzing residuals helps diminishing such extreme results. Anyhow, 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. Both group mean residuals are zero, suggesting the absence of differential prediction. The coexistence of overprediction and underprediction remains undetected.

Methodological Conclusions. With regard to future metaanalysis 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, that is, the region where the persons are located who are most likely to be affected by differential prediction.

Final Conclusion

The present meta-analysis summarizes the results of 130 samples and 493,048 students. It shows that admission tests underpredict women's academic performance and overpredict men's academic performance to a small but consistent extent. Consequently, women with the same academic potential as men would be less likely to be admitted if admission tests are used as the sole criterion for admission. This conclusion holds true for older as well as for newer tests.

The underprediction of women's academic performance is not related to predictor differences or criterion differences. Particularly undergraduate admission tests are more prone to over- and underprediction effects than graduate tests. These results point to gender differences of undergraduate students in studying habits and motivational factors. Future research should build on these results and focus on sex differences in non-cognitive factors of undergraduate students rather than on test or criterion differences.

Finally, the validity of admission tests is well documented and alternative ways of student selection which are both reasonable and feasible are scarce. Until the reasons for the underprediction of women's academic performance are better understood, we recommend using admission tests only in combination with HGPA or UGPA.

4 Sex Differences in Secondary School Success: Why Female Students Perform Better

[Geschlechtsunterschiede im Abiturerfolg: Warum Frauen besser abschneiden]

Franziska Fischer,

Johannes Schult &

Benedikt Hell

Universität Konstanz

Post-print manuscript; This article may not exactly replicate the final version published in the European Journal of Psychology of Education. It is not the copy of record. The final publication is available at http://link.springer.com/article/10.1007%2Fs10212-012-0127-4.

Fischer, F., Schult, J. & Hell, B. (2013). Sex differences in secondary school success: Why female students perform better. *European Journal of Psychology of Education*, *28*(2), 529-543. doi: 10.1007/s10212-012-0127-4

4.1 Abstract

School success is closely linked to intelligence but also to non-cognitive factors such as achievement motivation. The present study examines which non-cognitive factors predict secondary school grades and looks at reasons why female students tend to outperform their male counterparts. A sample of 554 German freshman students provide measures of general intelligence, achievement motivation, science course choice, self-efficacy, self-perceived academic achievement, and test anxiety. Results show that achievement motivation, self-perceived academic achievement, and sex significantly contribute to the final secondary school success above intelligence. Females' advantage in final secondary school grades becomes even larger after controlling for general intelligence. This advantage can be explained by females' higher achievement motivation. Showing more compensatory effort as well as self-control and taking more pride in their own productivity helps females to outperform their male counterparts at secondary school.

Keywords: School achievement, sex differences, achievement motivation, personality, intelligence

4.2 Introduction

Girls outperform boys at school (Ellis et al., 2008, p. 278; Deary, Strand, Smith, & Fernandes 2007; Downey & Vogt Yuan, 2005). This difference is at odds with the fact that females do not surpass males in general intelligence (Halpern, 2000, p. 128). The nature and underlying mechanisms of these discrepancies are unclear. Especially sex differences in final secondary school grades are rarely investigated though final secondary school success determines educational perspectives (e.g., in gaining acceptance onto selective universities, courses of study, and scholarships).

The most effective predictor of academic success is general intelligence (Gottfredson, 2003; Kuncel, Hezlett, & Ones, 2004; Sternberg, Grigorenko, & Bundy, 2001). Among school students the correlation between general intelligence and school achievement is about $r = .50$ (Furnham, Monsen, & Ahmetoglu, 2009; Jensen, 1980, p. 319; Laidra, Pullmann, & Allik, 2007; Lu, Weber, Spinath, & Shi, 2011), nominating intelligence as a promising reason of sex differences in academic success. Reviews, however, indicate that there are no inherent sex differences in general intelligence. There are just small differences in some specific ability domains (Blakemore, Berenbaum, & Liben, 2009, p. 94; Hedges & Nowell, 1995; Hyde, 2005). Consequently, it takes no wonder that intelligence fails to account for sex differences in school achievement (Freudenthaler, Spinath, & Neubauer, 2008). Only specific abilities can explain a small amount of sex differences in school success (Calvin, Fernandes, Smith, Visscher, & Deary, 2010). Still, what is the main reason for girls outperforming boys at school? The answer probably lies in non-cognitive factors (Furnham & Monsen, 2009; Petrides, Chamorro-Premuzic, Frederickson, & Furnham, 2005).

The influence of sex differences in non-cognitive qualities on school success is not fully explored (e.g., De Fruyt, Van Leeuwen, De Bolle, & De Clercq, 2008; Steinmayr & Spinath, 2008). School anxiety plays a role in determining academic success in boys, whereas work avoidance is an issue for girls (Freudenthaler et al., 2008). Extraversion is associated with better German and English grades for girls but not for boys (B. Spinath,

Freudenthaler, & Neubauer, 2010). These findings, however, are all related to middle school success. The present study expands the scope by focusing on final secondary school performance and, in particular, by investigating sex-specific influences of non-cognitive factors on final secondary school grades.

More girls than boys graduate from secondary school (USA: Heckman & LaFontaine, 2007; Germany: Hannover & Kessels, 2011; Statistisches Bundesamt, 2010). Given this fact, it is interesting to look at the extent of females' superior final secondary school grades after controlling for intelligence. Females' grades at middle school are significantly higher than expected by their intelligence scores (Duckworth & Seligman, 2006; Steinmayr & Spinath, 2008; Lewis & Hoover, 1987). Comparable results are found with regard to college admission tests which are quite similar to intelligence measures (Frey & Detterman, 2004; Jackson & Rushton, 2006). With the same test result women earn .12 points higher grades than men on average (based on a 4-point scale; Young & Kobrin, 2001). The present study explores if these findings can be transferred to final secondary school achievement by determining the extent of girls' advantage in final secondary school grades above intelligence.

There is an apparent sex difference in grades after controlling for intelligence. This raises the question which non-cognitive qualities help female students to outperform their male classmates. Factors improving the prediction accuracy of school achievement do not necessarily explain females' better school grades. Related research findings exist for the college context. Female students show different study habits compared to males (e.g., Robbins, Lauver, Davis, Langley, & Carlstrom, 2004). These habits include a tendency to devote more effort to academic work, show more class attendance as well as greater academic motivation (e.g., Zwick 2002, p. 151). By accounting for such variables, sex differences in college grades can be reduced (Stricker et al., 1993; Veldman, 1968). The present study tests whether non-cognitive factors can explain females' better secondary school success (when school grades are controlled for intelligence).

Potential Non-Cognitive Factors in Accounting for Success in Secondary School of Girls and Boys

In the following, findings regarding non-cognitive variables which contribute to a successful performance at school or in college are presented. Sex-specific differences are highlighted in order to identify the most promising variables to a) *optimize the prediction* of sex-specific academic success, and b) *explain sex-related differences* in academic success.

"Whether or not someone performs a task depends not only on 'can one' but also on 'will one'" (Mouw & Khanna 1993, p. 334; see also Cronbach, 1949). Accordingly, achievement motivation is positively correlated with school (Schuler & Prochaska, 2000) and college success (Robbins et al., 2004). In turn, achievement motivation is associated with conscientiousness (Schuler & Prochaska, 2000). Being conscientious strongly correlates with school grades (De Raad & Schouwenburg, 1996; Poropat, 2009), college performance (Trapmann, Hell, Hirn, & Schuler, 2007), and job success (Barrick, Mount, & Judge, 2001). Some researchers even consider achievement motivation a facet of conscientiousness (Costa & McCrae, 1998). Being diligent and highly motivated has beneficial effects on study habits which can explain females' better college grades above intelligence (Stricker et al., 1993; Veldman, 1968). Sex differences in achievement motivation exist for different aspects of this domain. They affect individual compensatory effort, pride in productivity, and self-control with female students scoring higher (Schuler & Prochaska, 2000).

Academic performance is positively correlated with self-efficacy (r = .38; Multon, Brown, & Lent, 1991). Males tend to show higher levels of self-efficacy compared to their female counterparts (Hinz, Schumacher, Albani, Schmid, & Brähler, 2006).

Self-perceived academic achievement is also positively associated with academic performance (Shen & Pedulla, 2000; Steinmayr & Spinath, 2007). Achievements in school are affected by self-perceived abilities even when cognitive ability is controlled

for (B. Spinath, Spinath, Harlaar, & Plomin, 2006; F. M. Spinath, Spinath, & Plomin, 2008; Steinmayr & Spinath, 2009). Longitudinal data support this relationship (Chamorro-Premuzic, Harlaar, Greven, & Plomin, 2010). Perceived ability is higher for male students particularly in mathematics (F. M. Spinath et al., 2008).

In contrast to self-efficacy and self-perceived academic achievement, test anxiety is associated with poor performance in school. Meta-analyses yield negative relations to academic performance of about -.21 (Seipp, 1991; Hembree, 1988). In regard to test performance the association is more pronounced for boys than for girls (McCarthy & Goffin, 2005). This is surprising as females typically show more anxiety than males (Feingold, 1994; Zeidner, 1998, p. 262).

Another way to explain females' superior grades in college is their tendency to enroll in courses which are less stringent and have more lenient grading systems. Female students show a tendency of avoiding particularly science classes (van Langen, Rekers-Mombarg, & Dekkers, 2006). Corrections for differences in grading standards and course selection patterns can significantly reduce differences in college grades (Elliott & Strenta, 1988; Hewitt & Goldman, 1975). Given these insights, the relationship between course choice and school achievement deserves a closer look in order to expand our knowledge concerning secondary school success.

Study Objectives

The present study investigates determinants of males' and females' secondary school success. At first, we explore sex differences in the impact of general intelligence, achievement motivation, science course choice, self-efficacy, self-perceived academic achievement, and test anxiety on final secondary school achievement. Secondly, we estimate female students' advantage in final secondary school grades after controlling for intelligence. Finally, we look into course choice, achievement motivation and further

non-cognitive factors that may explain females' superior grades after controlling for intelligence.

4.3 Method

Sample and Procedure

Freshmen students from two universities in Southern Germany were asked to participate voluntarily in this study. The students were matriculated in economics, business administration, or science. About 35% of the total freshmen population participated (that is 670 students; 317 males; 353 females). The majority (> 90%) of the participants stems from fields of study with no tangible admission restrictions. Therefore, there should be no considerable restriction of variance of final secondary school grades and intelligence.

The students participated in their free time. Within one session they first took part in an intelligence test that was administered in a group setting. Afterwards they filled in several questionnaires. Non-native speakers of the German language and participants who refused to indicate their final secondary school grades were excluded from the analyses. Only those candidates who scored better than what would be expected by guessing on all scales of the intelligence test were included for further evaluation. This process of elimination led to a final sample size of 264 male and 290 female students. The age of the participants ranged from 18 to 46 years (M = 20.26, Mdn = 20.00, SD = 1.83). Cases with further item non-response were excluded pairwise.

Measures

Academic performance. Participants reported their final secondary school GPA (the German Abiturnote). The German scoring system was reversed to obtain scores on a 4-point grading scale, with 4 representing the best possible result.

Intelligence. Candidates' cognitive ability was assessed using scales of a well-established German intelligence measure, the Intelligenz-Struktur-Test 2000 R (IST 2000 R; Liepmann, Beauducel, Brocke, & Amthauer, 2007). The IST 2000 R comprises three verbal, three numerical, and three figural reasoning tasks. For economical reasons, we administered two tests of each category. We selected those tests that yield the most reliable results. Verbal intelligence was measured using tests requiring the finding of verbal analogies and similarities. Numerical intelligence was measured using the tests numerical calculations and number series. Figural intelligence was determined using tests relating to figure selection and cubes. The scores achieved in the six reasoning tests were added up to a general intelligence measure.

Factor analysis of all cognitive test scores yielded a clear one-factor solution. The first unrotated principal component had an eigenvalue of 2.63 and accounted for 43.82% of the variance. The internal consistency of the intelligence scale was $\alpha = .90$.

Achievement motivation. Participants' level of achievement motivation was assessed using parts of the German equivalent of the Achievement Motivation Inventory (Schuler, Thornton, Frintrup, & Mueller-Hanson, 2004; German version: Leistungs-motivationsinventar; Schuler, Prochaska, & Frintrup, 2001). We chose those three subscales on which female candidates score significantly higher than their male counterparts (Schuler & Prochaska, 2000). The scales are compensatory effort, pride in productivity, and self-control. Candidates were asked to rank the applicability of statements to their own situation (e.g., "It makes me proud and happy to have mastered a difficult task."). The scale ranged from 1 (very inaccurate) to 5 (very accurate). Each

candidate's scores pertaining to the 15 items were aggregated to an overall score. The internal consistency of the scale was α = .82.

Science course choice. Participants reported which courses they had chosen during their last two years at secondary school. For further analysis we focused on students' science courses: mathematics, physics, chemistry, and biology. The variable science course choice was coded 1 when participants had taken at least two science subjects. Otherwise the variable was coded 0.

Self-efficacy. The German version of the General Self-Efficacy Scale (Skala zur Allgemeinen Selbstwirksamkeitserwartung; Schwarzer & Jerusalem, 1999) was used to measure perceived self-efficacy. Ten statements (e.g., "I can always manage to solve difficult problems if I try hard enough.") had to be rated on a 5-point scale. The choices ranged from 1 (not at all true) to 5 (exactly true). Candidates' responses were added up to an overall self-efficacy score. Cronbach's Alpha was α = .81.

Self-perceived academic achievement. Participants were asked to indicate their expected first year GPA at university. In the same way as with school GPA, the expected first year GPA was afterwards reversed to a score from 1 to 4, with 1 representing the lowest possible grade.

Test anxiety. Test anxiety was assessed by means of a new and shortened version of the German Test Anxiety Inventory (Hodapp, 1991), the Prüfungsangstfragebogen (Hodapp, Rohrmann, & Ringeisen, 2011). The short version comprises four scales: emotionality, worry, interference, and lack of confidence. Each scale consists of 5 items. Participants were asked to rank how they feel and what they think during test situations on a 4-point scale. The scale ranged from *almost never* to *almost always*. Responses to statements pertaining to lack of self-confidence were reversed before computing an individual's sum score. The internal consistency was α = .86.

Data Analysis

Three preliminary analyses were performed. Firstly, we determined correlations between the measures and age to identify possible age-related effects. Secondly, we tested for sex differences in the predictors and the criterion to compare the sample characteristics with previous findings. Thirdly, we explored the intercorrelations between the measures for both sexes. This was done to check the expected association between non-cognitive factors and school success and to get clear about overlapping constructs.

Our main analysis covered three examinations. In a first step we explored sex differences regarding the impact of different factors on final secondary school GPA. Therefore, a hierarchical multiple regression analysis was performed. General intelligence and non-cognitive variables were entered in one block (step 1), followed by the interactions between every predictor and sex (also entered in one block, step 2).

In a second step, we analyzed to what extent female students reach better final secondary school grades than their male counterparts, after controlling for general intelligence. We regressed final secondary school GPA on general intelligence. Afterwards we averaged the residuals for male and female students separately (i.e., we determined the amount of differential prediction).

Finally, we examined which personality and motivational variables explain females' higher final secondary grades above intelligence. Separate regression analyses for each explanatory variable were performed with respect to school success regressed on intelligence. For every regression model the mean residuals for male and female students were determined.

In the present study we used the term predictor in the technical sense of the word relating to regression analysis. Effectively, the predictors were measured shortly after the criteria. Given that the used intelligence and personality variables are stable traits, the results should not be affected by the measurement-sequence.

4.4 Results

Preparatory Analyses

Correlations with age. Results showed a significant correlation between age and final secondary school GPA (r = -.20, p < .01), achievement motivation (r = -.11, p < .05), self-efficacy (r = .10, p < .05), and self-perceived academic achievement (r = -.14, p < .01). Since age-related effects are not in the focus of the present study we regressed all predictor measures as well as school success on participants' age (like Freudenthaler et al., 2008 and B. Spinath et al., 2010). The resulting residuals were used in all subsequent data analyses.

Sex differences in predictors and criterion. Males scored significantly higher with regard to self-efficacy and self-perceived academic achievement. Female candidates showed significantly higher achievement motivation and test anxiety. The effects were small to moderate (see Table 4.1).

Table 4.1

Descriptive statistics (corrected for age) by sex and statistics for sex differences

	Males		Females			
	M	*SD*	*M*	*SD*	*p*	*d*
Final secondary school GPA	-.03	.58	.02	.54	.29	-.09
General intelligence	3.54	13.55	-3.23	13.15	< .01	.49
Achievement motivation	-2.41	7.98	2.24	7.77	< .01	-.58
Science course choice	.04	.49	-.03	.50	.09	.15
Self-efficacy	.55	3.54	-.52	3.97	< .01	.28
Self-perceived academic achievement	.06	.43	-.06	.40	< .01	.29
Test anxiety	-1.61	7.97	1.53	9.29	< .01	-.36

Note. Male students: 238 < n < 264; Female students: 250 < n < 290.

Contrary to the initial hypothesis there was no sex difference in science course choice, female students had no advantage in final secondary school GPA, and male students scored better in general intelligence (d = .49; see Table 4.1).

Correlations between measures. General intelligence was moderately correlated with secondary school success in male students (r = .43) and female students alike (r = .37). This difference was not significant (Fisher's z < 1.96). In male students school success was also significantly associated with all other predictors. In female students no significant association could be established with regard to science course choice and self-efficacy (see Table 4.2). Comparisons of the correlation between secondary school success and the aforementioned predictors showed no significant sex-related differences (all Fisher's z < 1.96).

Analysis of the intercorrelation between the predictor variables showed a similar pattern for males and females. In both sexes, test anxiety was negatively associated with general intelligence, self-efficacy, and self-perceived academic achievement. General intelligence and self-efficacy both correlated considerably with self-perceived academic achievement (cf. Table 4.2). Achievement motivation was significantly associated with self-efficacy and, in the female sample, also with self-perceived academic achievement. The only substantial difference between sexes referred to the strength of the link between achievement motivation and science course choice (Fisher's z = 2.41).

Table 4.2

Correlations of all measures for males (above diagonal) and females (below diagonal)

	Final secondary school GPA	General intelligence	Achievement motivation	Science course choice	Self-efficacy	Self-perceived academic achievement	Test anxiety
Final secondary school GPA	-	.43**	.32**	.18**	.15*	.31**	-.22**
General intelligence	.37**	-	-.11	.09	.08	.16*	-.28**
Achievement motivation	.23**	-.11	-	.12	.26**	.06	-.03
Science course choice	.04	.10	-.10	-	.06	.07	-.05
Self-efficacy	.01	.04	.25**	.05	-	.27**	-.38**
Self-perceived academic achievement	.40**	.24**	.17**	.07	.18**	-	-.14*
Test anxiety	-.23**	-.19**	-.02	.07	-.41**	-.31**	-

Note. Male students: $220 < n < 264$; Female students: $232 < n < 289$.

* $p < .05$ ** $p < .01$.

Multiple Regression Analyses Pertaining to Male and Female Students

Table 4.3 shows the results of the hierarchical multiple regression analysis. Candidates' final secondary school GPA was used as the dependent variable. General intelligence, achievement motivation, science course choice, self-efficacy, self-perceived academic achievement, and test anxiety were used as predictors (step 1), as well as their interactions with sex (step 2). The final model explained 35% of the variance in secondary school achievement ($F(13, 409) = 16.93$, $p < .001$). The interactions accounted for 1.2% of the variance and did not improve the model significantly ($F(6, 409) = 1.23$, $p = .289$). General intelligence appeared to constitute the strongest predictor, followed by achievement motivation, self-perceived academic achievement, and sex. Science course choice, self-efficacy, and test-anxiety failed to contribute significantly to the prediction of final secondary school GPA.

Table 4.3

Predicting final secondary school GPA by intelligence, non-cognitive factors (step 1) and their interactions with sex (step 2) with N = 423

Predictors	R^2	*B*	β	*t*	*sr*
Step 1	.338				
General intelligence		.017	.425	6.89**	.274
Achievement motivation		.024	.354	5.74**	.229
Science course choice		.067	.060	1.03	.041
Self-efficacy		-.004	-.026	-.36	-.014
Self-perceived academic achievement		.272	.208	3.53**	.141
Test anxiety		-.005	-.084	-1.20	-.048
Female		.125	.112	2.54*	.101
Step 2	.350				
General intelligence * Female		-.005	-.089	-1.50	-.060
Achievement motivation * Female		-.006	-.064	-1.04	-.042
Science course choice * Female		-.075	-.048	-.83	-.033
Self-efficacy * Female		-.020	-.103	-1.46	-.058
Self-perceived academic achievement * Female		.081	.043	.71	.028
Test anxiety * Female		-.007	-.082	-1.16	-.046

Note. sr = semi-partial correlation.

* $p < .05$ ** $p < .01$.

Sex-Specific Differential Prediction of Academic Success Through General Intelligence

In order to determine the extent of differential prediction, a candidate's final secondary school GPA was regressed on general intelligence. After accounting for general intelligence, the difference in secondary school achievement between male and female students amounted to .15 grade points (on average). Female students earned grades which were .07 grade points higher than predicted by their general intelligence

level (mean residual = .07, 95% CI = [.02, .13]). In contrast, male students achieved grades which were .08 grade points lower than predicted based on their general level of intelligence (mean residual = -.08, 95% CI = [-.14, -.02]).

Sex-Specific Differential Prediction of Academic Success Through Explanatory Variables

Regressions for every explanatory variable were performed. The corresponding sex-specific residuals are presented in Table 4.4. Only achievement motivation successfully reduced differential prediction. After controlling for achievement motivation, the mean residuals for male and female students no longer showed significant deviation from zero. In contrast, science course choice, self-efficacy, self-perceived academic achievement, and test anxiety failed to eliminate sex-related differences in final secondary school success.

Table 4.4

Mean residuals using explanatory variables as predictors for final secondary school GPA[a]

	Males			Females		
Explanatory Variables	*N*	*Res*	95% CI	*N*	*Res*	95% CI
Achievement motivation	247	-.03	[-.09, .03]	265	.02	[-.03, .08]
Science course choice	263	-.08	[-.15, -.02]	289	.08	[.02, .13]
Self-efficacy	239	-.08	[-.15, -.02]	253	.08	[.02, .14]
Self-perceived academic achievement	244	-.12	[-.18, -.06]	265	.11	[.05, .17]
Test anxiety	237	-.09	[-.16, -.03]	250	.09	[.02, .15]

Note. *Res* = mean residual; CI = confidence interval.

[a]GPA is controlled for intelligence

4.5 Discussion

Main Results

Final secondary school performance of boys and girls strongly depends on general intelligence. This finding is in line with existing empirical evidence on middle school success (Freudenthaler et al., 2008; B. Spinath et al., 2010) and academic achievement in general (Gottfredson, 2003; Kuncel et al., 2004; Sternberg et al., 2001; Hell et al., 2007). Our results underline the influence of personality variables on school grades beside general intelligence (Day, Hanson, Maltby, Proctor, & Wood, 2010; Leeson, Ciarrochi, & Heaven, 2008). Achievement motivation, self-perceived academic achievement, as well as sex significantly contribute to secondary school success. The significance of self-perceived academic achievement is consistent with recent findings about middle school students (B. Spinath et al., 2006; F. M. Spinath et al., 2008; Steinmayr & Spinath, 2009). The study further exposes that science course choice, self-efficacy and test anxiety do not contribute to the prediction in any considerable way. Still, the constructs behind these explanatory variables are partially overlapping (see preparatory analysis). Using the mentioned variables as single predictors may yield different results. The absence of an effect of self-efficacy on academic success may be due to the more general level of the applied self-efficacy scale. A more specific scale measuring *academic* self-efficacy might have revealed the expected impact. Academic self-efficacy describes a more direct self-evaluation of one's ability for success in the academic environment. Studies applying academic self-efficacy items could show the impact of self-efficacy on school success (Grigorenko et al., 2009) and college success (Robbins et al., 2004; Chemers, Hu, & Garcia, 2001).

None of the interactions between the personality variables and sex reaches significance. Also their semi-partial correlations indicate very small effects. In other words, achievement motivation, science course choice, self-efficacy, self-perceived academic achievement, and test anxiety show no sex differences in the prediction of final secondary school success.

After controlling for intelligence, female students' advantage in final secondary school GPA amounts to .15 grade points (on a 4-point grading scale). In other words, females earn grades which are .15 grade points higher than that of their male counterparts with the same general intelligence. These findings point to the magnitude of the effect of girls' non-cognitive qualities on college success. The results add strength to previous research which found that sex-related differences in school performance cannot be explained by general intelligence (Freudenthaler et al., 2008; B. Spinath et al., 2010). In fact, controlling for cognitive ability even enlarges the gender gap.

As reported, achievement motivation and self-perceived academic achievement improve the prediction accuracy of secondary school success above intelligence. Anyhow, only achievement motivation can explain females' superior final secondary school success. Female students show more compensatory effort as well as self-control and they take more pride in their own productivity. These characteristics explain females' superior grades.

It is worth noting that controlling for self-perceived academic achievement even increases the gender gap in grades, though self-perceived academic achievement improves the prediction accuracy. The present results provide support for the claim that adjusting for course choice does not reduce differential prediction of academic success (Leonard & Jiang, 1999; Stricker et al., 1993).

Limitations and Future Research

The females in our sample show only marginally higher final secondary school GPAs and have lower intelligence scores than the male candidates. These surprising results could be a consequence of sampling. Our sample does not cover the entire spectrum of secondary school graduates and omits those who decided not to pursue tertiary education. Still, participants perform only 6% better on the intelligence measure than the normative sample of secondary school students and their mean final secondary

school GPA was only 8% above the national average. As a further explanation our sample might comprise more gifted males than females because science subjects are less attractive to gifted female students than to gifted males (Gavin & Reis, 2005). This premise finds some support in the separate analysis of economics and business administration students (N_{Males} = 136, $N_{Females}$ = 157). In this subsample, female students have higher secondary school grades than their male colleagues (d = -.21; M_{Males} = -.27, SD_{Males} = .48, $M_{Females}$ = -.16, $SD_{Females}$ = .51) and the comparative advantage of male students in terms of general ability is smaller (d = .40; M_{Males} = -1.32, SD_{Males} = 12.72, $M_{Females}$ = -6.44, $SD_{Females}$ = 12.71).

The remaining sex difference with regard to general intelligence mirrors the findings of Colom and Lynn (2004) as well as of Irwing and Lynn (2005). Both studies note a small general intelligence difference in favor of male candidates. Other studies using the IST 2000 R find similar effects (B. Spinath et al., 2010; Steinmayr, Beauducel, & Spinath, 2010). It appears that the gap between women and men in general intelligence can be attributed to the nature of the applied intelligence measure. Another plausible explanation relates to the characteristics of intelligence distribution for the sexes. Males have been found to be represented more frequently at the extreme ends of the intelligence distribution (Hedges & Nowell, 1995; Strand, Deary, & Smith, 2006). Final secondary school graduates tend to lie above average on the intelligence distribution. This fact may explain the advantage of males in general intelligence within our sample. Further research in this field would benefit from using a representative sample of students, ideally, one that captures the natural spectrum that exists at the point of graduation from secondary school.

The cross-sectional design of this study prevents the determination of causal links. Future research could also address the interesting question how personality and motivational factors influence school performance. The relationship between conscientiousness and college performance, for example, is mediated by increased academic effort (Noftle & Robins, 2007). Already existing multidimensional models

about the interplay of motivation, skill, and performance may be of relevance to future research (e.g., Eccles & Wigfield, 2002).

General intelligence, coupled with the employed personality and motivational variables, cannot explain all of the variance in secondary school achievement. Future studies in this field may wish to investigate the impact of additional variables including academic self-efficacy (Grigorenko et al., 2009), hopefulness (Day et al., 2010), and self-discipline (Duckworth & Seligman, 2005).

Finally, from an applied point of view, it would be of interest to further investigate how achievement motivation can be increased in males. Moreover, it is worth examining why the higher grades and the higher achievement motivation do not constitute in terms of females' long-term career performance. Our findings suggest that achievement motivation is a promising lead when it comes to identifying non-cognitive factors that are related to sex differences in secondary school success.

5 Unterschätzung der Studienleistungen von Frauen durch Studierfähigkeitstests: Erklärbar durch Persönlichkeitseigenschaften?

Franziska Fischer,

Johannes Schult &

Benedikt Hell

Universität Konstanz

Pre-Print Manuskript. Dieser Beitrag wurde in leicht abgeänderter Form bei einer deutschsprachigen Zeitschrift eingereicht und ist in dieser Form noch nicht zur Veröffentlichung angenommen.

Fischer, F., Schult, J. & Hell, B. (2012). *Unterschätzung der Studienleistungen von Frauen durch Studierfähigkeitstests: Erklärbar durch Persönlichkeitseigenschaften?* Manuskript eingereicht zur Publikation.

5.1 Zusammenfassung

In einem längsschnittlichen Studiendesign wird für zwei deutschsprachige Studierfähigkeitstests untersucht, ob diese die Studienleistungen von Frauen unterschätzen und wenn ja, ob Persönlichkeitseigenschaften diese Unterschätzung erklären können. Die Datenbasis liefern 356 Studienanfänger in Wirtschaftswissenschaften und 269 Studienanfänger in Naturwissenschaften. Neben den Noten des ersten Studienjahrs werden die Leistungen im Studierfähigkeitstest, der Abiturerfolg sowie die Leistungsmotivation, die Selbstdisziplin und die Allgemeine Selbstwirksamkeit der Studienanfänger erhoben. Im Ergebnis unterschätzen die eingesetzten fachspezifischen Studierfähigkeitstests die Studienleistungen von Frauen, insbesondere im oberen Leistungsbereich, deutlich. Demgegenüber unterschätzt die Abiturnote die Studienleistungen von Männern im obersten Leistungsbereich. Studierfähigkeitstest und Abiturerfolg zusammen liefern sowohl die valideste Vorhersage als auch die geringste geschlechtsspezifische Über- bzw. Unterschätzung der Studiennoten. Facetten der Leistungsmotivation und Selbstdisziplin können die Unterschätzung von Frauen durch die Studierfähigkeitstests teilweise erklären, wohingegen die Allgemeine Selbstwirksamkeit sich nicht auf die Unterschätzung der Studienleistungen auswirkt.

Schlüsselwörter: Studierfähigkeitstest, Schulnoten, Studienerfolg, Studierendenauswahl, Fairness, Bias, Studienberatung

5.2 Einleitung

Die Zahl der Studienanfänger an deutschen Universitäten ist so hoch wie nie zuvor (Statistisches Bundesamt, 2012). Infolge der Aussetzung der Wehrpflicht und der Umstellung auf das achtjährige Gymnasium in den großen Bundesländern werden die Kapazitäten der Hochschulen ausgereizt. Gleichzeitig bricht etwa jeder vierte Studienanfänger sein Studium ab oder wechselt das Studienfach (Heublein, Hutzsch, Schreiber, Sommer & Besuch, 2009; Statistisches Bundesamt, 2011). Vor diesem Hintergrund formulierte der Wissenschaftsrat (2004) folgendes Ziel: „Studierwillige müssen weit mehr als bisher ein Studium aufnehmen, das ihren Fähigkeiten und Neigungen in besonderem Maße entspricht“ (S. 33). Die Beratungsangebote für Studieninteressierte wurden daraufhin verbessert und den Hochschulen wurde mehr Autonomie bei der Vergabe ihrer Studienplätze zugesprochen (vgl. Siebtes Gesetz zur Änderung des Hochschulrahmengesetzes). Seitdem ist der Einsatz von Studierfähigkeitstests in Deutschland deutlich gestiegen.

Ein wesentlicher Bestandteil der gesetzlichen Neuregelung besteht darin, dass Hochschulen auch fachspezifische Studierfähigkeitstests bei der Zulassung einsetzen dürfen. Neben zahlreichen medizinischen Fakultäten nutzen vor allem private Hochschulen sowie Fachhochschulen diese Möglichkeit, letztere insbesondere bei der Zulassung zu wirtschaftswissenschaftlichen Studiengängen (Zimmerhofer & Trost, 2008). Aber auch für das Fach Psychologie wurden bereits Testverfahren erprobt und erfolgreich eingesetzt (z. B. Formazin et al., 2011). Studierfähigkeitstests werden als Chance gesehen, Leistungs- und Vorkenntnisniveau der Studienanfänger zu homogenisieren, eine fachspezifische Eignung der Bewerber zu sichern und letztendlich die Abbruchquoten zu senken (Moosbrugger, Jonkisz & Fucks, 2006).

Auch im Beratungskontext werden Studierfähigkeitstests eingesetzt. Sowohl die Bundesagentur für Arbeit als auch Self-Assessment-Angebote, wie beispielsweise was-studiere-ich.de (Hell, Päßler & Schuler, 2009), verwenden Studierfähigkeitstests, um Studieninteressenten ein Feedback darüber zu geben, inwiefern ihr angestrebtes

Studienfeld ihren Fähigkeiten entspricht (Kubinger et al., 2012; Psychologischer Dienst, 2011).

Die zunehmende Bedeutung von Studierfähigkeitstests im Auswahl- und Beratungskontext liegt in deren guter Vorhersagekraft begründet. Metaanalysen bestätigen Studierfähigkeitstests eine prognostische Validität zwischen .41 und .59 (nach Korrektur für die Reliabilität der Studiennoten und/oder Variabilitätseinschränkungen; Hell et al., 2007; Kuncel & Hezlett, 2007; Patterson & Mattern, 2011). Studierfähigkeitstests sind zudem über die Abiturnote hinaus inkrementell valide (Hell et al., 2008; Kobrin et al., 2008; Trost, Klieme & Nauels, 1997). Gleichzeitig lassen US-amerikanische Forschungsbefunde jedoch Zweifel aufkommen, ob Studierfähigkeitstests fair gegenüber weiblichen Studienbewerbern sind (Young & Kobrin, 2001).

Wann genau ein Test als unfair zu bezeichnen ist, wurde in der Vergangenheit kontrovers diskutiert (Meade & Fetzer, 2009; Wottawa & Amelang, 1980). Hierbei wurden sowohl soziale, als auch psychometrische Aspekte betrachtet (SIOP, 2003). Die Standards for Educational and Psychological Testing (AERA, APA & NCME, 1999, S. 74-80) unterscheiden zusammenfassend drei psychometrische Anforderungen an einen fairen Test. Ein Test darf für bestimmte Teilnehmergruppen a) kein DIF, b) keine *differenziellen Validitäten* und c) keine *differenzielle Prädiktion* aufzeigen. Dabei ist es wichtig differenzielle Prädiktion von differenzieller Validität abzugrenzen (Young & Kobrin, 2001). Differenzielle Prädiktion liegt vor, wenn sich die Regressionsgeraden, welche den Zusammenhang zwischen Test und Kriterium abbilden, für bestimmte Teilgruppen (z. B. Männer und Frauen) signifikant unterscheiden (Meade & Fetzer, 2009, S. 740). Die Kriteriumsleistung der Gruppe, deren Regressionsgerade über der Regressionsgeraden der anderen Gruppe liegt, wird dabei durch den Test unterschätzt.

Amerikanische Studierfähigkeitstests sind ausführlich auf ihre geschlechts-spezifische differenzielle Prognose hin untersucht (z. B. Patterson et al., 2009). Fischer, Schult und Hell (2012a) fassen diese Forschungsarbeiten in einer Metaanalyse zusammen und finden eine generalisierbare Unterschätzung der Studienleistungen von

Frauen. Bei gleichen Testergebnissen erreichen Frauen Studiennoten, die im Schnitt um .24 Notenpunkte besser sind als die ihrer männlichen Kommilitonen. Ob diese Befunde auch auf deutschsprachige Studierfähigkeitstests übertragbar sind, ist offen, da deutsche Tests diesbezüglich bisher unzureichend untersucht wurden. Die wenigen Befunde, die vorliegen, deuten auf eine geringe Unterschätzung der Studienleistungen von Frauen hin (Nauels & Meyer, 1997; Dlugosch, 2005).

Mögliche Ursachen für die Unterschätzung der Studienleistungen von Frauen durch Studierfähigkeitstests

Lange wurde die Ursache für das schlechtere Abschneiden von Frauen in Studierfähigkeitstests bestimmten Testinhalten zugeschrieben. Älteren Testverfahren wird angelastet, dass sich ihre Items häufiger mit typisch männlichen Domänen befassen, wodurch Frauen benachteiligt werden (Zwick, 2002, S. 151). Studierfähigkeitstests jüngeren Datums weisen jedoch keine Auffälligkeiten in den Iteminhalten auf und unterschätzen dennoch die Studienleistungen von Frauen (Patterson et al., 2009).

Ungünstig auf die Leistungen von Frauen in Testsituationen kann sich ebenfalls der Einfluss von *stereotype threat* auswirken (Steele, 1997; Walton & Spencer, 2009). Wird während der Testsituation das Stereotyp aktiviert, dass Frauen schlechtere Leistungen in mathematischen Bereichen erzielen, verschlechtert sich deren Leistung in mathematischen Aufgabengruppen (Spencer et al., 1999).

Ein mehrfach empirisch überprüfter Ansatz sieht die Ursache für die prädiktive Verzerrung im unterschiedlichen Kurswahlverhalten von Frauen und Männern (Alon & Gelbgiser, 2011; Berry & Sackett, 2009). Frauen wählen tendenziell einfachere Vertiefungsrichtungen oder Studienschwerpunkte, erhalten dort bessere Noten und werden letztendlich von Studierfähigkeitstests unterschätzt, da das Notenkriterium verzerrt ist (Elliott & Strenta, 1988; Hewitt & Goldman, 1975). Obwohl andere

empirische Studien keinen Einfluss des Kurswahlverhaltens auf die Unterschätzung der Studienleistungen von Frauen nachweisen können (Leonard & Jiang, 1999; Stricker et al., 1993), ist dies einer der populärsten Erklärungsansätze.

Persönlichkeitseigenschaften und damit zusammenhängendes Studienverhalten werden ebenfalls als Ursache für die Unterschätzung von Frauen durch Studierfähigkeitstests gesehen (Sackett et al., 2008). Studentinnen sind demnach engagierter, nehmen regelmäßiger am Unterricht teil und sind motivierter, wodurch sie bei gleichen kognitiven Kapazitäten bessere Noten erzielen als ihre männlichen Kommilitonen (Zwick, 2002, S. 151). Leistungsmotivation und Selbstdisziplin scheinen vielversprechende Konstrukte zu sein, zumal sie positiv mit Studienerfolg korreliert sind und bei Frauen häufig stärker ausgeprägt sind (Robbins et al., 2004). Zudem konnte gezeigt werden, dass Selbstdisziplin die Unterschätzung der Schulleistungen von Schülerinnen der achten Klasse durch standardisierte Tests erklären kann (Duckworth & Seligman, 2006). Ebenso kann Leistungsmotivation die besseren Abiturnoten von Frauen unter Kontrolle der kognitiven Fähigkeiten aufklären (Fischer, Schult & Hell, 2012b). Im Zusammenhang mit Studienleistungen fehlen hierzu jedoch empirische Untersuchungen.

Bedeutendster nicht-kognitiver Faktor in der Vorhersage von Studienerfolg ist Selbstwirksamkeit (Richardson, Abraham & Bond, 2012; Robbins et al., 2004). Es werden gewöhnlich Korrelationen um $r = .59$ erreicht. Über einen möglichen Zusammenhang mit der Über- bzw. Unterschätzung von Studienleistungen ist bislang nichts bekannt.

Fragestellungen und Vorgehen

Ziel der vorliegenden Studie ist es, für zwei deutschsprachige Studierfähigkeitstests zu untersuchen, ob diese in ihrer Vorhersage unfair sind gegenüber Frauen (im Sinne der differenziellen Prädiktion) und ob diese Unfairness

verschiedene Leistungsbereiche gleich stark betrifft. Zur Klärung der Forschungsfrage wird eine möglichst unselektierte Bewerberkohorte rekrutiert, die ein breites Leistungsspektrum abdeckt, so dass erstens Varianzeinschränkungen die Aussagekraft der Ergebnisse nicht mindern und zweitens post hoc verschiedene Selektionsquoten analysiert werden können.

In der Praxis wird aus rechtlichen und praktischen Gründen bei der Studienzulassung meist auf die Abiturnote zurückgegriffen, weshalb ergänzend untersucht wird, ob auch die Abiturnote und eine Kombination aus Abiturnote und der Leistung im Studierfähigkeitstest den Studienerfolg verzerrt vorhersagen.

Sollte sich bestätigen, dass die Studierfähigkeitstests zu einer unfairen Vorhersage des Studienerfolgs führen, wird geprüft, ob die zusätzliche Berücksichtigung von Persönlichkeitseigenschaften der Verzerrung entgegen wirken bzw. diese potenziell erklären kann. Im Speziellen wird getestet, ob die Hinzunahme von Leistungsmotivation, Selbstdisziplin sowie Allgemeiner Selbstwirksamkeit zu einer fairen Prognose des Studienerfolgs für Männer und Frauen führt.

5.3 Methode

Stichprobe und Durchführung

Zu Beginn des Wintersemesters 2010/2011 nahmen 671 Studienanfänger zweier baden-württembergischer Universitäten freiwillig an der Untersuchung teil[9]. Die Probanden hatten sich entweder für ein wirtschaftswissenschaftliches oder ein naturwissenschaftliches Bachelor- oder Lehramtsstudium immatrikuliert und waren zum Zeitpunkt der Testung im Durchschnitt 20.3 Jahre alt (*SD* = 1.8). Es wurden gezielt Studienanfänger rekrutiert, da diese in den ersten beiden Semestern ihres Studiums

[9] Die Ergebnisse sind Teil einer umfangreichen Längsschnittstudie. Weitere Befunde zu verwandten Fragestellungen sind zu finden in Fischer et al. (2012b) und Schult, Fischer und Hell (2012).

dieselben Grundlagenprüfungen absolvieren müssen und somit auf Kurswahleffekte zurückzuführende Verzerrungen des Kriteriums ausgeschlossen werden können. Die Teilnehmer absolvierten einen ihrem gewählten Studienfeld entsprechenden Studierfähigkeitstest und füllten Fragebögen aus. Es wurden ausschließlich weibliche Testleiter eingesetzt, um die Aktivierung des Stereotyps, dass Frauen schlechtere Leistungen in mathematischen Bereichen erzielen, und einen daraus resultierenden leistungsminderenden Effekt zu vermeiden.

Ein Jahr später wurden mit Hilfe der Zentralen Datenschutzstelle der baden-württembergischen Universitäten und den entsprechenden Prüfungsämtern die erbrachten Studienleistungen ausgelesen. Es lagen letztendlich Testleistungen und Studiennoten von 356 Wirtschaftswissenschaftlern (53% Frauen) und 269 Naturwissenschaftlern (52% Frauen) vor. Die Stichprobe der Naturwissenschaftler umfasste hierbei 73 Biologiestudenten, 62 Studenten der Ernährungs- bzw. Lebensmittelwissenschaften, 45 Physikstudenten, 44 Chemiestudenten, 24 Life-Science-Studenten und 21 Studenten anderer naturwissenschaftlicher Fächer. Abgesehen von Life-Science unterlag die tatsächliche Zulassung zu den in der Strichprobe vertretenen Studienfächern keinen besonderen Restriktionen.

Prädiktorvariablen

Studierfähigkeitstest. Es wurden die als *Studienfeldbezogene Beratungstests* konzipierten Studierfähigkeitstests der Bundesagentur für Arbeit für die Fachrichtungen Wirtschaftswissenschaften und Naturwissenschaften eingesetzt. Beide Tests wurden von der ITB Consulting GmbH entwickelt. Sehr ähnliche Testvarianten werden von verschiedenen Hochschulen zur Auswahl ihrer Studierenden verwendet. Die beiden Studierfähigkeitstests umfassen 40 Multiple-Choice-Aufgaben mit jeweils einer richtigen Antwortalternative. Die Verfahren bestehen aus jeweils zwei unterschiedlichen Aufgabengruppen bei einer Gesamtdurchführungszeit von 120 bzw. 105 Minuten. Der

wirtschaftswissenschaftliche Test ist in die Testteile *Modellanalyse* sowie *Diagramme und Tabellen* untergliedert. Im Aufgabenteil *Modellanalyse* müssen die Probanden komplexe, formalisierte Funktionssysteme verstehen und auf konkrete Fragestellungen anwenden. Im Testteil *Diagramme und Tabellen* sind Grafiken, Schaubilder und Tabellen aus dem wirtschaftlichen Bereich zu analysieren und zu interpretieren. Der naturwissenschaftliche Test enthält ebenfalls eine Aufgabengruppe mit *Diagrammen und Tabellen*, welche sehr ähnlich aufgebaut ist wie im wirtschaftswissenschaftlichen Test, jedoch stammen die hier dargebotenen Informationen aus dem naturwissenschaftlichen Bereich. Der zweite Testteil trägt den Namen *Naturwissenschaftliches Grundverständnis*. Hierin wird das Strukturieren verbal vermittelter naturwissenschaftlicher Informationen gefordert.

Abiturerfolg. Die Studienanfänger berichteten ihre Gesamtnote im Abitur (oder die Note eines äquivalenten Schulabschlusses, der die Aufnahme eines Studiums an einer deutschen Universität ermöglicht). Ausländische Noten wurden in das deutsche Notensystem überführt.

Leistungsmotivation. Die Leistungsmotivation der Probanden wurde mit Hilfe des Leistungsmotivationsinventars (Schuler et al., 2001) gemessen. Es wurden die Subskalen Kompensatorische Anstrengung, Leistungsstolz und Selbstkontrolle verwendet. Diese Facetten sind bei Frauen stärker ausgeprägt als bei Männern, und sie weisen eine positive Validität für Studienleistungen auf (Schuler & Prochaska, 2000). Diese Kombination (Mittelwertsdifferenzen und positive Validität) macht die genannten Facetten zu potentiellen erklärenden Variablen für die erwarteten Verzerrungseffekte. Insgesamt umfassten die drei Subskalen 15 Items. Zu jedem Item gaben die Probanden auf einer fünfstufigen Skala an, in welchem Ausmaß dieses auf sie persönlich zutrifft (1 = trifft nicht zu, 5 = trifft zu). Für die Gesamtstichprobe ergab sich ein Cronbachs Alpha von .82 über alle Items hinweg.

Selbstdisziplin. Selbstdisziplin, als Unterfacette der Dimension Gewissenhaftigkeit des Fünf-Faktoren-Modells (Costa & McCrae, 1992), wurde mit Items aus dem

International Personality Item Pool erhoben (Goldberg et al., 2006). Die deutsche Übersetzung nach Zumdick (2007) umfasst 10 Items, zu denen die Probanden auf einer fünfstufigen Skala angaben, wie stark die einzelnen Aussagen auf sie zutreffen (1 = trifft nicht zu, 5 = trifft zu). Für die Gesamtstichprobe zeigte sich eine interne Konsistenz von $\alpha = .86$.

Allgemeine Selbstwirksamkeit. Es wurde die Skala zur Allgemeinen Selbstwirksamkeitserwartung von Schwarzer und Jerusalem (1999) eingesetzt. Die Skala misst mit 10 Items die optimistische Kompetenzerwartung (z. B. "Schwierigkeiten sehe ich gelassen entgegen, weil ich meinen Fähigkeiten immer vertrauen kann."). Auf einer vierstufigen Skala gaben die Probanden an, wie zutreffend die einzelnen Aussagen für sie sind (1 = stimmt nicht, 4 = stimmt genau). Cronbachs Alpha betrug $\alpha = .81$.

Kriteriumsvariable Studienerfolg

Als Maß für den Studienerfolg wurde die Durchschnittsnote der im ersten Studienjahr erbrachten Studienleistungen berechnet. Die Studienleistungen der Wirtschaftswissenschaftler wurden zunächst innerhalb der beiden Universitäts-Stichproben standardisiert und daraufhin zusammengefasst. Anschließend wurden die Studiennoten zur besseren Veranschaulichung wieder linear in eine fünfstufige Notenskala transformiert.

Datenanalyse

Die Daten wurden getrennt für wirtschaftswissenschaftliche und naturwissenschaftliche Studiengänge ausgewertet. Deskriptive Statistiken und Interkorrelationen aller Variablen wurden bestimmt und alle Variablen wurden auf geschlechtsspezifische Unterschiede getestet. Fehlende Werte wurden paarweise ausgeschlossen.

Um die Studierfähigkeitstests auf geschlechtsspezifische differenzielle Prädiktion zu untersuchen wurde das Vorgehen von Cleary (1968) angewandt. Dieser Ansatz sieht vor, den Zusammenhang zwischen den Leistungen im Test und dem Studienerfolg in geschlechtsspezifischen Regressionsgeraden abzubilden und diese auf Unterschiede in ihren Achsenabschnitten und Steigungsparametern zu testen. Für den Fall, dass sich die Achsenabschnitte und/oder Steigungsparameter signifikant unterscheiden, ist der Test nach Cleary (1968) als unfair zu bezeichnen. Dieses Vorgehen wird in der englischsprachigen Literatur auch als *Cleary approach* bezeichnet (z. B. Linn, 1973; Meade & Tonidandel, 2010). Um einen Vergleich der Regressionsordinaten im relevanten mittleren Wertebereich zu ermöglichen wurden die Testleistungen zunächst standardisiert (vgl. Schmidt & Hunter, 1982). Anschließend wurden die Unterschiede in den Achsenabschnitten und Steigungsparametern mit Hilfe einer moderierten multiplen Regression getestet, wobei der Haupteffekt Geschlecht dem Unterschied in den Achsenabschnitten und die Interaktion Studierfähigkeitstest x Geschlecht dem Unterschied in den Steigungsparametern entspricht.

Eine Aussage bezüglich der Über- bzw. Unterschätzung eines Geschlechts für alle Leistungsbereiche kann getroffen werden, wenn sich nur die Achsenabschnitte unterscheiden. Das Geschlecht mit dem größeren Achsenabschnitt wird unterschätzt, da es im Mittel bessere Studienleistungen erzielt als es eine gemeinsame Regressionsgerade vorhersagt. Schneiden sich die geschlechtsspezifischen Regressionsgeraden, liegen unterschiedliche Befunde für verschiedene Leistungsbereiche vor. Dann ist es hilfreich, den Verlauf der Regressionsgeraden anhand eines Schaubildes zu betrachten bzw. geschlechtsspezifische Residuen (s. folgender Abschnitt) für verschiedene Leistungsbereiche zu berechnen.

Als Alternative zum *Cleary approach* stellte Lawshe (1983) ein einfacheres Verfahren zur Analyse der differenziellen Prädiktion vor. Zu diesem Zweck wird eine *gemeinsame* Regressionsgerade bestimmt, welche den Zusammenhang zwischen Testleistung und Studienerfolg abbildet. Ausgehend von dieser Regressionsgeraden

werden die mittleren geschlechtsspezifischen Residuen berechnet (z. B. Bridgeman et al., 2008; Mattern et al., 2008). Der Vorteil dieses Vorgehens besteht darin, dass die mittleren Residuen ein direktes und sehr anschauliches Maß für die Über- bzw. Unterschätzung der Studienleistung in Notenpunkten bieten. Sie geben an, um wie viele Notenpunkte eine Gruppe im Mittel besser bzw. schlechter im Studium abschneidet als es eine gemeinsame Regressionsgerade vorhersagen würde. Ein weiterer Vorteil dieser Methode besteht darin, dass geschlechtsspezifische Residuen auch für spezifische Leistungsbereiche betrachtet werden können.

Um die Frage zu beantworten, ob verschiedene Bereiche gleich stark von der differenziellen Prädiktion betroffen sind, wurde berechnet, wie groß die mittleren geschlechtsspezifischen Residuen für die jeweils 10%, 25%, 35% und 75% Besten des Studierfähigkeitstests bzw. Vorhersagemodells ausfallen. Der Unterschied in den mittleren geschlechtsspezifischen Residuen wurde für jede Selektionsquote zudem als Effektstärke ausgedrückt (vgl. Lawshe, 1983; Stricker et al., 1993).

Kurz zusammengefasst wurden zunächst die geschlechtsspezifischen Regressionsgeraden auf Unterschiede in ihren Achsenabschnitten und Steigungsparametern getestet. Anschließend wurden für verschiedene Leistungsbereiche die mittleren geschlechtsspezifischen Residuen berechnet, die auf einer *gemeinsamen* Regressionsgeraden basieren. Diese Berechnungen wurden unabhängig voneinander für die Prädiktoren Studierfähigkeitstest und Abiturnote durchgeführt.

Neben der Leistung im Studierfähigkeitstest und dem Abiturerfolg als Einzelprädiktoren wurden Vorhersagemodelle aus einer Kombination von Testleistung und jeweils einer Erklärungsvariable gebildet. Mit dieser Methode wurde überprüft, ob unter Berücksichtigung von bestimmten Persönlichkeitsmerkmalen (bzw. der Abiturnote) die Unfairness in der Vorhersage des Studienerfolgs durch Studierfähigkeitstests aufgeklärt werden kann. Es entstanden die Vorhersagemodelle Studierfähigkeitstest & Abitur, Studierfähigkeitstest & Leistungsmotivation, Studierfähigkeitstest & Selbstdisziplin sowie Studierfähigkeitstest & Selbstwirksamkeit.

Die Gewichtung der einzelnen Modellkomponenten erfolgte hierbei gemäß den optimalen β-Gewichten aus den entsprechenden multiplen Regressionen mit Studienerfolg als abhängiger Variable. Jedes einzelne Vorhersagemodell wurde analog zum oben beschriebenen Vorgehen auf seine geschlechtsspezifische differenzielle Prädiktion getestet (Vergleich der geschlechtsspezifischen Regressionsgeraden und Bestimmung der geschlechtsspezifischen Residuen).

5.4 Ergebnisse

Deskriptives, geschlechtsspezifische Unterschiede und Interkorrelationen

Tabelle 5.1 zeigt für die Stichprobe Wirtschaftswissenschaften die deskriptiven Statistiken sowie die Interkorrelationen aller Variablen getrennt für Männer und Frauen. Tabelle 5.2 zeigt analog die Ergebnisse für die Stichprobe Naturwissenschaften. Zusätzlich werden in beiden Tabellen Effektstärken für das Ausmaß der Geschlechtsunterschiede in den einzelnen Variablen berichtet.

Männliche Studienanfänger schneiden sowohl im wirtschaftswissenschaftlichen als auch im naturwissenschaftlichen Studierfähigkeitstest deutlich besser ab als weibliche Studienanfänger (d_{WIWI} = 0.64 bzw. d_{NAT} = 1.01), erzielen jedoch keine besseren Studiennoten. Frauen zeigen signifikant mehr Leistungsmotivation und Selbstdisziplin als Männer.

In beiden Stichproben korreliert der Studienerfolg am höchsten mit der Abiturnote und den Leistungen im Studierfähigkeitstest. Leistungsmotivation ist erwartungsgemäß hoch korreliert mit Selbstdisziplin.

Tabelle 5.1

Interkorrelationen, Mittelwerte, Standardabweichungen und Effektstärken der Variablen für die Stichprobe Wirtschaftswissenschaften

Variable	1	2	3	4	5	6	Männer		Frauen		p	Cohens d
							M	*SD*	*M*	*SD*		
1. Studiennoten	-	-.41**	.55**	-.14	-.07	.07	2.92	0.86	3.07	0.89	.11	-.17
2. SFT	-.49**	-	-.35**	-.01	-.09	-.02	20.02	6.70	15.84	5.28	< .01	.69
3. Abitur	.60**	-.37**	-	-.32**	-.15	-.02	2.45	0.49	2.34	0.53	< .05	.22
4. Leistungsmotivation	-.07	-.08	-.21*	-	.69**	.28**	22.71	8.01	27.67	7.48	< .01	-.64
5. Selbstdisziplin	-.03	-.11	-.20*	.68**	-	.26**	29.07	7.01	33.84	6.54	< .01	-.70
6. Selbstwirksamkeit	.02	.08	-.04	.35**	.32**	-	29.23	4.05	28.99	4.02	.61	.06

Anmerkungen. Interkorrelationen für Männer ($139 < n < 169$) werden oberhalb der Diagonalen dargestellt und Interkorrelationen für Frauen ($142 < n < 187$) unterhalb der Diagonalen. Positive Effektstärken bedeuten eine höhere Merkmalsausprägung zu Gunsten der Männer. SFT = Studierfähigkeitstest.

* $p < .05$. ** $p < .01$.

Tabelle 5.2

Interkorrelationen, Mittelwerte, Standardabweichungen und Effektstärken der Variablen für die Stichprobe Naturwissenschaften

Variable	1	2	3	4	5	6	Männer		Frauen		p	Cohens d
							M	*SD*	*M*	*SD*		
1. Studiennoten	-	-.26**	.47**	-.24**	-.15	.04	2.67	0.85	2.69	0.82	.86	.02
2. SFT	-.38**	-	-.44**	.02	.02	-.04	27.36	5.13	22.13	5.20	< .01	1.01
3. Abitur	.41**	-.45**	-	-.31**	-.26**	.02	1.93	0.55	1.93	0.52	.94	.00
4. Leistungsmotivation	-.13	-.04	-.33**	-	.73**	.11	23.18	7.55	27.09	7.64	< .01	-.51
5. Selbstdisziplin	-.21*	-.05	-.27**	.79**	-	.24**	29.57	6.38	32.30	7.24	< .01	-.40
6. Selbstwirksamkeit	-.07	.11	-.14	.14	.19*	-	29.94	3.39	27.83	4.02	< .01	.57

Anmerkungen. Interkorrelationen für Männer ($113 < n < 129$) werden oberhalb der Diagonalen dargestellt und Interkorrelationen für Frauen ($122 < n < 140$) unterhalb der Diagonalen. Positive Effektstärken bedeuten eine höhere Merkmalsausprägung zu Gunsten der Männer. SFT = Studierfähigkeitstest.

* $p < .05$. ** $p < .01$.

Genderfairness von Studierfähigkeitstests und Abitur in der Prädiktion des Studienerfolgs

Die Ergebnisse der moderierten multiplen Regressionen zeigt Tabelle 5.3, die Ergebnisse der Residuen-Analysen für verschiedene Selektionsquoten sind in Tabelle 5.4 dargestellt.

Stichprobe Wirtschaftswissenschaften. Die Vorhersage des Studienerfolgs basierend auf den Leistungen im Studierfähigkeitstest ist nach Clearys Definition (1968) als unfair zu bezeichnen. Abbildung 5.1 zeigt, dass sich die geschlechtsspezifischen Regressionsgeraden im unteren Leistungsbereich kaum unterscheiden, während im oberen Leistungsbereich die Regressionsgerade der Frauen deutlich über der Regressionsgerade der Männer liegt, was für eine Unterschätzung der Frauen im oberen Leistungsbereich spricht. Die Analysen der geschlechtsspezifischen Residuen bestätigen dieses Bild. Für Selektionsquoten von 35% bis 10% beträgt der Geschlechtsunterschied in den Residuen mindestens $d = .40$ (s. Tabelle 5.4).

Für die Vorhersage basierend auf dem Abiturerfolg ergibt sich ein anderes Bild (vgl. Abbildung 5.1). Die Regressionsgerade der Männer liegt signifikant über der Regressionsgeraden der Frauen (vgl. Tabelle 5.3). Der Abstand der Regressionsgeraden nimmt dabei mit steigender Abiturleistung ab. Die Analyse der Residuen bestätigt eine Unterschätzung der Männer insbesondere im unteren und obersten Leistungsbereich (s. Tabelle 5.4). Das Vorhersagemodell Studierfähigkeitstest & Abitur zeigt für alle Selektionsquoten eine faire Vorhersage.

Tabelle 5.3

Ergebnisse der moderierten multiplen Regressionen für verschiedene Vorhersagemodelle mit Studienerfolg als abhängiger Variable

Stich-probe	Vorhersagemodell	Test der Steigungsparameter			Test der Achsenabschnitte			Bewertung nach Cleary[a]
	Prädiktor(en)	*F*	*df*	*p*	*F*	*df*	*p*	
WIWI	SFT	4.82	1; 354	< .05	2.09	1; 354	.15	unfair
	Abitur	0.17	1; 331	.68	8.06	1; 331	< .01	unfair
	SFT & Abitur	1.89	1; 331	.17	0.64	1; 331	.43	fair
	SFT & Leistungsmotivation	5.63	1; 289	< .05	1.52	1; 289	.22	unfair
	SFT & Selbstdisziplin	5.88	1; 288	< .05	1.38	1; 288	.24	unfair
NAT	SFT	0.83	1; 267	.36	5.30	1; 267	< .05	unfair
	Abitur	0.21	1; 261	.65	0.30	1; 261	.58	fair
	SFT & Abitur	0.01	1; 261	.94	0.71	1; 261	.40	fair
	SFT & Leistungsmotivation	0.00	1; 262	.97	2.44	1; 262	.12	fair
	SFT & Selbstdisziplin	1.06	1; 263	.31	2.56	1; 263	.11	fair

Anmerkungen. WIWI = Wirtschaftswissenschaften, NAT = Naturwissenschaften; SFT = Studierfähigkeitstest.

[a] Nach Cleary (1968) ist ein Test unfair, wenn die gruppenspezifischen Regressionsgeraden unterschiedliche Steigungen und/oder unterschiedliche Ordinatenabschnitte aufweisen.

* $p < .05$. ** $p < .01$.

Tabelle 5.4

Geschlechtsspezifische mittlere Residuen für verschiedene Selektionsquoten und Vorhersagemodelle

			Selektionsquote													
			100%			75%			35%			25%			10%	
Stichprobe	Vorhersagemodell	Geschlecht	*N*	*Res*	*d*	*n*	*Res*	*d*	*n*	*Res*	*d*	*n*	*Res*	*d*	*n*	*Res*	*d*
WIWI	SFT	M	169	0.06	.13	149	0.04	.16	91	0.17	**.60**	62	0.15	**.47**	26	0.27	**1.07**
		F	187	-0.06		132	-0.10		56	-0.36		24	-0.32		4	-0.80	
	Abitur	M	158	-0.13	**-.31**	113	-0.16	**-.26**	50	-0.08	-.10	43	-0.07	-.11	13	-0.16	**-.38**
		F	175	0.12		138	0.10		68	0.02		54	0.04		27	0.15	
	SFT & Abitur	M	158	0.04	.09	120	-0.01	.02	58	0.05	.20	44	0.06	.11	15	0.11	-.04
		F	175	-0.03		130	-0.01		56	-0.11		43	-0.03		18	0.15	
	SFT & Leistungsmotivation	M	141	0.06	.13	110	0.04	.15	59	0.20	**.53**	50	0.29	**.68**	22	0.35	**.98**
		F	150	-0.06		106	-0.09		43	-0.28		22	-0.31		7	-0.53	
	SFT & Selbstdisziplin	M	142	0.06	.12	115	0.03	.14	60	0.14	**.46**	50	0.23	**.58**	23	0.28	**.81**
		F	148	-0.05		103	-0.10		41	-0.28		23	-0.30		6	-0.46	
NAT	SFT	M	129	0.11	**.25**	117	0.12	**.37**	69	0.08	**.21**	54	0.15	**.30**	21	-0.01	**.37**
		F	140	-0.10		89	-0.18		17	-0.12		10	-0.09		4	-0.32	
	Abitur	M	126	-0.03	-.07	89	-0.06	-.11	46	-0.01	.01	36	0.01	-.05	11	-0.15	**-.34**
		F	137	0.02		107	0.02		39	-0.02		29	0.05		13	0.11	
	SFT & Abitur	M	126	0.04	.10	99	0.05	.10	53	-0.03	.07	43	0.00	-.04	16	-0.06	-.07
		F	137	-0.04		99	-0.02		39	-0.08		22	0.03		10	-0.02	
	SFT & Leistungsmotivation	M	127	0.07	.19	107	0.07	.19	55	-0.03	.13	46	0.03	**.22**	19	-0.01	**.22**
		F	137	-0.07		91	-0.08		37	-0.13		20	-0.19		7	0.15	
	SFT & Selbstdisziplin	M	129	0.07	.19	109	0.08	**.24**	59	0.09	**.29**	46	0.13	**.47**	17	0.01	**.60**
		F	136	-0.07		90	-0.10		34	-0.14		20	-0.25		9	-0.45	

Anmerkungen. Die geschlechtsspezifischen Residuen beschreiben die mittlere Überschätzung (positives Vorzeichen) bzw. Unterschätzung (negatives Vorzeichen) der Studienleistung in Notenpunkten basierend auf einer gemeinsamen Regressionsgeraden für das jeweilige Vorhersagemodell. Effektstärken > .20 sind fettgedruckt. WIWI = Wirtschaftswissenschaften, NAT = Naturwissenschaften; SFT = Studierfähigkeitstest; F = Frauen, M = Männer.

Stichprobe Naturwissenschaften. Auch in der Stichprobe Naturwissenschaften ist die Vorhersage des Studienerfolgs basierend auf den Leistungen im Studierfähigkeitstest unfair. Die geschlechtsspezifischen Regressionsgeraden verlaufen hierbei ähnlich wie die Geraden der Wirtschaftswissenschaftler (vgl. Abbildung 5.1). Im Gegensatz zur Stichprobe Wirtschaftswissenschaften zeigt sich jedoch über *alle* Selektionsquoten hinweg eine relativ konstante Unterschätzung der Studienleistungen von Frauen (s. Tabelle 5.4).

Für den Prädiktor Abitur unterscheiden sich die Regressionsgeraden nicht (vgl. Tabelle 5.3 und Abbildung 5.1). Lediglich im obersten Leistungsbereich werden die Studienleistungen von Männern unterschätzt (s. Tabelle 5.4). Analog zu den Wirtschaftswissenschaftlern liefert das Vorhersagemodell Studierfähigkeitstest & Abitur auch für die Naturwissenschaftler eine faire Vorhersage über alle Selektionsquoten hinweg.

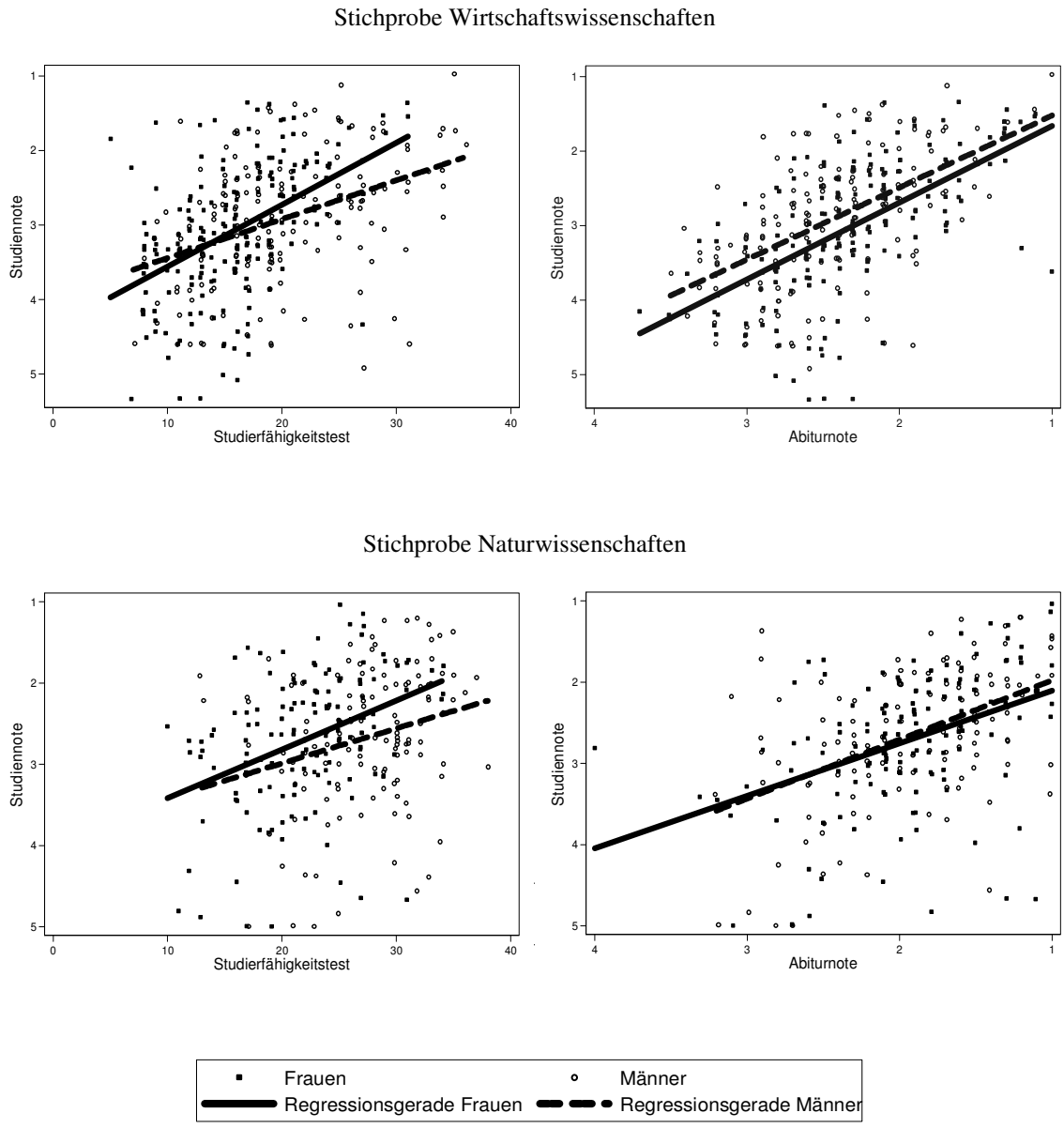

Abbildung 5.1. Die Schaubilder zeigen den Studienerfolg in Abhängigkeit von der Leistung im Studierfähigkeitstest bzw. dem Abiturerfolg, getrennt für die Stichprobe Wirtschaftswissenschaften und die Stichprobe Naturwissenschaften. Zusätzlich sind die jeweiligen geschlechtsspezifischen Regressionsgeraden dargestellt. Unterschiede in den Regressionsgeraden deuten auf eine unfaire Vorhersage hin.

Genderfairness von Studierfähigkeitstests & Persönlichkeitseigenschaften in der Prädiktion des Studienerfolgs

Tabelle 5.5 beschreibt die β-Gewichte der einzelnen Prädiktoren aus den multiplen Regressionen, welche als Gewichte für die Bildung der im Folgenden analysierten Vorhersagemodelle (und den bereits erwähnten Vorhersagemodellen Studierfähigkeitstest & Abitur) verwendet wurden. Zusätzlich werden die Validitäten der Vorhersagemodelle sowie der Einzelprädiktoren Studierfähigkeitstest und Abiturerfolg angegeben. In beiden Stichproben liefert das Modell Studierfähigkeitstest & Abitur die valideste Vorhersage. Ebenso erweisen sich Leistungsmotivation und Selbstdisziplin jeweils in Kombination mit den Leistungen im Studierfähigkeitstest als signifikante Prädiktoren. Entgegen den Erwartungen ist Allgemeine Selbstwirksamkeit in Kombination mit den Leistungen im Studierfähigkeitstest kein signifikanter Prädiktor für den Studienerfolg. Dementsprechend ist kein Einfluss auf die Fairness des Studierfähigkeitstests zu erwarten, weshalb diese Prädiktorkombination für die Fairness-Analysen im Folgenden nicht weiter berücksichtigt wird.

Innerhalb der Stichprobe Wirtschaftswissenschaften liefert keine Kombination aus Studierfähigkeitstest und Persönlichkeitseigenschaft eine faire Vorhersage. Innerhalb der naturwissenschaftlichen Stichprobe fällt die Vorhersage für die Modelle Studierfähigkeitstest & Leistungsmotivation sowie Studierfähigkeitstest & Selbstdisziplin nach Cleary (1968) fair aus (vgl. Tabelle 5.3). Betrachtet man jedoch die geschlechtsspezifischen Residuen, wird deutlich, dass sich die Unterschätzung der Frauen durch die Hinzunahme von Leistungsmotivation und Selbstdisziplin zwar verringert, jedoch vor allem im oberen Leistungsbereich nicht vollständig verschwindet (vgl. Tabelle 5.4).

Tabelle 5.5

Kennwerte multipler Regressionen verschiedener Vorhersagemodelle mit Studienerfolg als abhängige Variable

Stichprobe	Prädiktor(en)	*N*	Prädiktor 1		Prädiktor 2		Validität
			β	*t*	*β*	*t*	
WIWI	SFT	356					-.44
	Abitur	333					.56
	SFT & Abitur	333	-.28	-6.30**	.48	10.58**	.63
	SFT & Leistungsmotivation	291	-.45	-8.43**	-.15	-2.82**	.45
	SFT & Selbstdisziplin	290	-.45	-8.44**	-.12	-2.25*	.45
	SFT & Selbstwirksamkeit	281	-.45	-8.35**	.07	1.25	.45
NAT	SFT	269					-.29
	Abitur	263					.44
	SFT & Abitur	263	-.17	-2.84**	.38	6.31**	.47
	SFT & Leistungsmotivation	264	-.33	-5.66**	-.22	-3.70**	.37
	SFT & Selbstdisziplin	265	-.31	-5.31**	-.21	-3.59**	.35
	SFT & Selbstwirksamkeit	255	-.31	-5.03**	.02	0.39	.30

Anmerkungen. WIWI = Wirtschaftswissenschaften, NAT = Naturwissenschaften; SFT = Studierfähigkeitstest.
* $p < .05$. ** $p < .01$.

5.5 Diskussion

Die Ergebnisse der vorliegenden Längsschnittuntersuchung zeigen, dass die verwendeten Studierfähigkeitstests die Studienleistungen der Frauen unterschätzen. Frauen erzielen bei gleicher Testleistung in den Wirtschaftswissenschaften durchschnittlich um 0.12 Notenpunkte bessere Studienleistungen als ihre männlichen Kommilitonen. In den Naturwissenschaften sind es sogar 0.21 Notenpunkte. Diese

Unterschätzung liegt im Bereich der Befunde der Metaanalyse von Fischer et al., (2012a), die einen mittleren Unterschied von 0.24 Notenpunkten für überwiegend amerikanische Tests berichten.

Neu ist die Erkenntnis, dass die Unterschätzung innerhalb sehr strenger Selektionsquoten besonders hoch ausfällt. Bei einer Selektionsquote von 25% beträgt der Unterschied in den Studiennoten zwischen Männern und Frauen 0.47 Notenpunkte bei den Wirtschaftswissenschaftlern und 0.24 Notenpunkte bei den Naturwissenschaftlern. Diese Befunde bieten wichtige Erkenntnisse für die Frage der Hochschulzulassung, da Studierfähigkeitstests im Zulassungskontext vor allem dann eingesetzt werden, wenn eine strenge Auswahl vorgenommen werden muss, die Selektionsquote also sehr klein ist.

Während *High-School-Noten* die Studienleistungen von Frauen ebenfalls leicht unterschätzen (Fischer et al., 2012a), *überschätzt* die Abiturnote tendenziell die Studienleistungen der Frauen in Deutschland. Anders gesagt, die Abiturnote unterschätzt tendenziell die Studienleistungen von Männern. Dies gilt vor allem im obersten Leistungsbereich, nur in der Stichprobe Wirtschaftswissenschaften ist zusätzlich auch der untere Leistungsbereich betroffen. Diese Befunde machen deutlich, dass eine Zulassung in den untersuchten Fächern ausschließlich anhand der Abiturnote bei sehr strengen Selektionsquoten Männer benachteiligt.

Es stellt sich die Frage, warum nur in der wirtschaftswissenschaftlichen Stichprobe die Unterschätzung der Frauen durch den Test besonders stark den oberen Leistungsbereich betrifft und warum die Abiturnote auch im unteren Leistungsbereich Männer unterschätzt. Ein genauer Vergleich der beiden Stichproben deutet auf Selbstselektionsprozesse bei der Studienfachwahl hin. Die Abiturnoten von Studienanfängern in naturwissenschaftlichen Studiengängen (M = 1.93; SD = 0.53) sind bedeutend besser (d = 0.90) als die Abiturnoten von Studienanfängern in wirtschaftswissenschaftlichen Studiengängen (M = 2.40; SD = 0.51). Ähnliches gilt für die Leistungen im Studierfähigkeitstest und in den Studiennoten. Aus den Schaubildern

in Abbildung 5.1 wird dementsprechend deutlich, dass sich die Stichprobe der Wirtschaftswissenschaftler im Schnitt in einem niedrigeren Leistungsbereich bewegt als die Stichprobe der Naturwissenschaftler. Es kann daher spekuliert werden, dass sich ohne diese Selbstselektion die gefundenen Unterschiede in den Stichproben nicht zeigen würden, sondern die Unterschätzung der Frauen im oberen Leistungsbereich bei den Wirtschaftswissenschaftlern abnehmen und die Unterschätzung der Männer auch im unteren Leistungsbereich bei den Naturwissenschaftlern auftreten würde.

Analog zu bestehenden Befunden erweist sich die Kombination aus Studierfähigkeitstest und Abiturerfolg als valider Prädiktor (Hell et al., 2008; Kobrin et al., 2008). Mit unkorrigierten Korrelationen von .63 für die Wirtschaftswissenschaftler und .47 für die Naturwissenschaftler erzielt dieses Modell die valideste Vorhersage. Zudem zeigt das Vorhersagemodell über alle Selektionsquoten hinweg für beide Stichproben keine geschlechtsspezifische Über- oder Unterschätzung der Studienleistungen.

Auffallend ist, dass die Validitäten für die Vorhersage des Studienerfolgs in der naturwissenschaftlichen Stichprobe allgemein geringer ausfallen. Dies könnte an der heterogeneren Zusammensetzung der naturwissenschaftlichen Stichprobe liegen. Während die wirtschaftswissenschaftliche Stichprobe aus zwei homogenen Teilstichproben besteht, fließen in die naturwissenschaftliche Stichprobe mehr als sechs verschiedene Studienfächer ein.

Die vorliegende Studie zeigt, dass die eingesetzten deutschsprachigen Studierfähigkeitstests die Studienleistungen von Frauen unterschätzen, auch wenn keine Kurswahlmöglichkeiten innerhalb der Studienfächer bestehen. Dieses Ergebnis unterstützen die Befunde von Leonard und Jiang (1999), die ebenfalls keinen Einfluss des Kurswahlverhaltens auf die Unterschätzung der Frauen finden.

Persönlichkeitsmerkmale erklären indessen in bestimmten Konstellationen die geschlechtsspezifische Fairness der Studierfähigkeitstests. Durch die Berücksichtigung

von Leistungsmotivationsfacetten oder Selbstdisziplin verbessert sich die Fairness für naturwissenschaftliche Studienanfänger, jedoch zeigen die Ergebnisse der Residuen-Analysen, dass in den oberen Leistungsbereichen die Unfairness gegenüber Frauen bestehen bleibt. Die Unfairness scheint dabei etwas stärker mit der Leistungsmotivation als mit der Selbstdisziplin von Frauen zusammenzuhängen. Innerhalb der Gruppe der wirtschaftswissenschaftlichen Studienanfänger zeigt sich hingegen kein Effekt der Persönlichkeitsmerkmale. Während Leistungsmotivation und Selbstdisziplin im Schulkontext die besseren Leistungen von Frauen weitestgehend erklären können (vgl. Fischer et al., 2012b; Duckworth & Seligman, 2006), gelingt dies im Studienkontext nur eingeschränkt.

Entgegen den Erwartungen hängt die Allgemeine Selbstwirksamkeit nicht positiv mit dem Studienerfolg zusammen. Dieser Befund deutet darauf hin, dass das Konstrukt der Allgemeinen Selbstwirksamkeit zu unspezifisch ist, wenn es um die Vorhersage von Studienerfolg geht. Der akademische Bezug scheint entscheidend zu sein (Chemers et al., 2001; Robbins et al., 2004).

Keines der berücksichtigten Persönlichkeitsmerkmale kann die Unterschätzung der Studienleistungen von Frauen durch die eingesetzten Studierfähigkeitstests vollständig aufklären. Vielmehr scheint ein Zusammenspiel mit weiteren Faktoren für die Unterschätzung der Studienleistungen von Frauen verantwortlich zu sein (Meade & Fetzer, 2009; Meade & Tonidandel, 2010). Ein Ansatzpunkt für zukünftige Forschungsarbeiten könnte eine genauere Analyse des Abiturerfolgs der Frauen sein. Die Abiturnote misst offenbar etwas, das den Studienerfolg von Frauen eher überschätzt und damit in Verbindung mit den Leistungen im Studierfähigkeitstest zu einer genaueren Prognose führt. In diesem Zusammenhang ist es hilfreich zu betrachten, was die gewöhnlich weiblichen *Overachiever* in der Schule auszeichnet (Sparfeldt, Buch & Rost, 2010). Interessant wäre auch eine genauere Analyse des tatsächlichen Studienverhaltens von Männern und Frauen, sowohl im Bezug auf die Qualität (z. B. Lernstrategien) als auch die Quantität (z. B. Lerndauer). Ansätze liefert das multidimensionale Modell von

Eccles und Wigfield (2002). Es kann spekuliert werden, dass Leistungsmotivation und Selbstdisziplin nur unter weiter zu spezifizierenden Bedingungen zu einem günstigeren Studienverhalten führen.

Aus der vorliegenden Studie lassen sich folgende Schlussfolgerungen für die untersuchten Studierfähigkeitstests und Studienfächer ableiten: Bei sehr strengen Selektionsquoten ist sowohl der alleinige Einsatz von Abiturnoten als auch der alleinige Einsatz der Studierfähigkeitstests problematisch. Die Studierfähigkeitstests unterschätzen in diesem Leistungsbereich die Studienleistungen von Frauen, Abiturnoten diejenigen von Männern. Werden beide Merkmale bei der Zulassung kombiniert berücksichtigt, erhält man eine valide und gleichzeitig faire Vorhersage für Männer und Frauen. Weitere Forschungsarbeiten sollten anknüpfend an diese Befunde untersuchen, inwieweit diese Ergebnisse auf andere Instrumente und Testbedingungen übertragbar sind. Hierbei sollte die Generalisierbarkeit der gefundenen Ergebnisse auf a) unterschiedliche Konstruktbereiche der Tests (Intelligenzfacetten), b) mögliche Applikationsmethoden (Selbsttest vs. kontrollierte Testdurchführung), c) unterschiedliche Freiwilligkeitsgrade (freiwillige Teilnahme vs. Test als Zulassungsbedingung) und d) andere Studienfächer geprüft werden.

6 Gesamtdiskussion

Vorliegendes Kapitel gibt einen Überblick über die gewonnenen Ergebnisse dieser Dissertation und beantwortet die zu Beginn der Arbeit formulierten Forschungsfragen. Anknüpfend an die Ergebnisse werden Vorschläge für zukünftige Forschungstätigkeiten formuliert und es werden Implikationen sowohl für die Testentwicklung als auch für die Studienberatung und die Auswahl von Studierenden abgeleitet.

6.1 Zusammenfassung und Diskussion der Ergebnisse

Studierfähigkeitstests beeinflussen den Studien- und Berufsweg von vielen jungen Männern und Frauen. Trotz dieser großen Lenkungswirkung ist nicht hinreichend geklärt, ob Studierfähigkeitstests die Studienleistungen von Frauen unterschätzen und wenn ja, wie dieser Effekt erklärt werden kann. Die vorliegende Dissertation greift diese Fragen auf und liefert neue bedeutsame Erkenntnisse. Die wichtigsten Ziele und die Hauptergebnisse der durchgeführten Studien sind in Tabelle 6.1 zusammengefasst. Anschließend werden die in Abschnitt 2.8 aufgeworfenen Fragen anhand der gewonnenen Erkenntnisse studienübergreifend beantwortet und diskutiert.

Tabelle 6.1

Zusammenfassung der Ziele und Hauptergebnisse der empirischen Studien

Studie	Ziele	Hauptergebnisse
1 (Kapitel 3)	• Zusammenfassung des aktuellen internationalen Forschungsstands zur geschlechtsspezifischen differenziellen Prognose von Studienleistungen durch Studierfähigkeitstests (sowie von Test und Abiturnote) anhand metaanalytischer Techniken. • Identifikation von Moderatorvariablen, die mit der geschlechtsspezifischen differenziellen Prognose von Studierfähigkeitstests zusammenhängen.	• Die Studienleistungen von Frauen werden durch Studierfähigkeitstests unterschätzt (N = 493 048; d = .14). Bei gleicher Testleistung erzielen Frauen Studienleistungen, die um 0.24 Notenpunkte besser sind als die ihrer männlichen Kommilitonen. • Die Unterschätzung der Frauen fällt geringer aus, wenn bei der Vorhersage zusätzlich die Abiturnote berücksichtigt wird. • Testalter sowie geschlechtsspezifische Mittelwertsunterschiede im Test und in den Studiennoten sind keine signifikanten Moderatoren. • *Undergraduate-Tests* unterschätzen Frauen stärker als *Graduate-Tests* und mathematische Testteile unterschätzen Frauen stärker als verbale Testteile.
2 (Kapitel 4)	• Bestimmung von Persönlichkeitsmerkmalen, die nach Kontrolle der kognitiven Fähigkeiten den besseren Abiturerfolg von Frauen erklären.	• Die Leistungsmotivationsfacetten Kompensatorische Anstrengung, Leistungsstolz und Selbstkontrolle erklären die besseren Abiturleistungen von Frauen nach Kontrolle der kognitiven Fähigkeiten.
3 (Kapitel 5)	• Überprüfung zweier deutschsprachiger Studierfähigkeitstests hinsichtlich ihrer geschlechtsspezifischen differenziellen Prognose (sowie die differenzielle Prognose von Test und Abiturnote). • Untersuchung, ob die geschlechtsspezifische differenzielle Prognose alle Leistungsbereiche gleich stark betrifft. • Überprüfung, ob Persönlichkeitsunterschiede das Auftreten von differenzieller Prognose erklären können.	• Die eingesetzten deutschen fachspezifischen Studierfähigkeitstests unterschätzen die Studienleistungen von Frauen. • Die Unterschätzung von Frauen fällt im oberen Leistungsbereich besonders hoch aus. • Studierfähigkeitstest und Abiturerfolg zusammen führen zu einer genderfairen Vorhersage. • Leistungsmotivation und Selbstdisziplin können die Unterschätzung von Frauen durch Studierfähigkeitstests teilweise erklären.

I. Besteht eine generalisierbare Unterschätzung (im Sinne der differenziellen Prognose) von Frauen durch Studierfähigkeitstests?

Die Ergebnisse der Metaanalyse (Studie 1) bestätigen eine generalisierbare Unterschätzung der Studienleistungen von Frauen durch Studierfähigkeitstests. Bei gleichem Testergebnis erzielen Frauen im Schnitt um 0.24 Notenpunkte bessere Studienleistungen als Männer. Dieses Ausmaß an differenzieller Prädiktion ist vergleichbar mit den Ergebnissen für ethnische Minderheiten in den USA (Young & Kobrin, 2001).

In der vorliegenden Arbeit wurde darüber hinaus zum ersten Mal gezeigt, dass die geschlechtsspezifische differenzielle Prognose im oberen Leistungsbereich besonders stark ausfällt (Studie 3). Bei einer strengen Selektionsquote von z. B. 25% beträgt der Unterschied in den Studiennoten bis zu 0.47 Notenpunkte, d. h. Studierfähigkeitstests sind besonders dann unfair, wenn sehr streng selektiert wird. Bei einer Zulassungsquote von 25% würde dies für die untersuchten Stichproben folgendes bedeuten: obwohl sich unter den 25% besten Studierenden ohne Zulassungsbeschränkung 52% Frauen befinden, würde der Test als alleiniges Auswahlkriterium nur 28% Frauen zulassen. Hochgerechnet für 1 500 Studienplätze, die z. B. jährlich in Deutschland für das Fach Medizin vergeben werden, würde dies bedeuten, dass 360 Frauen zu Unrecht *keinen* Studienplatz erhalten würden, womit von einer hohen praktischen Bedeutsamkeit der gefundenen Effekte gesprochen werden kann.

II. Zeigt sich eine Unterschätzung der Frauen auch für deutschsprachige Studierfähigkeitstests?

In einem längsschnittlichen Studiendesign konnte für zwei deutschsprachige Studierfähigkeitstests gezeigt werden, dass auch diese Tests die Studienleistungen von Frauen unterschätzen (Studie 3). Bei gleicher Leistung im Studierfähigkeitstest erreichen Frauen um 0.12 bzw. 0.21 Notenpunkte bessere Studienleistungen als ihre männlichen

Kommilitonen. Die Unterschätzung der Studienleistungen von Frauen durch die eingesetzten deutschen Studierfähigkeitstests liegt damit im Bereich der Ergebnisse der Metanalyse.

III. Wirkt sich die Berücksichtigung der Abiturnote (neben den Leistungen im Studierfähigkeitstest) günstig auf die Vorhersage-Fairness aus – und wenn ja, warum?

Sowohl die Ergebnisse der Metaanalyse (Studie 1) als auch die Ergebnisse der Längsschnittstudie (Studie 3) belegen, dass die Vorhersage des Studienerfolgs durch einen Studierfähigkeitstest kombiniert mit der Abiturnote fairer ausfällt als durch einen Studierfähigkeitstest alleine. Frauen werden im Mittel um 0.08 Notenpunkte unterschätzt, wenn Studierfähigkeitstest und Abiturnote gemeinsam für die Vorhersage des Studienerfolgs herangezogen werden. Die beiden eingesetzten deutschen Studierfähigkeitstests führen in Verbindung mit den Abiturnoten sogar zu einer noch geringeren Unterschätzung der Frauen von 0.03 bzw. 0.04 Notenpunkten. Hier kann von einer fairen Vorhersage gesprochen werden.

Die Ergebnisse weisen darauf hin, dass die Abiturnote neben den kognitiven Fähigkeiten auch Persönlichkeitsmerkmale miterfasst, die entscheidend dafür sind, dass Frauen das vorhandene kognitive Potential besser in gute Studienleistungen umsetzen können. Studie 2 bestätigt, dass die Leistungsmotivationsfacetten Kompensatorische Anstrengung, Leistungsstolz und Selbstkontrolle bei Frauen stärker ausgeprägt sind und dass diese Facetten erklären können warum Frauen bessere Abiturleistungen erzielen als ihre männlichen Klassenkameraden (unter Kontrolle der kognitiven Fähigkeiten). Demnach scheint es wahrscheinlich, dass die Abiturnote zu einer fairen Vorhersage des Studienerfolgs führt, da sie die bei Frauen stärker ausgeprägte Leistungsmotivation mitberücksichtigt.

IV. Was sind die Ursachen für mögliche Unfairnessbefunde? Lassen sich die bestehenden Erklärungsansätze belegen und/oder durch neue Befunde erweitern?

Die Metaanalyse macht deutlich, dass ein besseres Abschneiden der Männer in Studierfähigkeitstests das Auftreten von differenzieller Prognose nicht moderiert. Bloße Mittelwertdifferenzen deuten nicht auf eine Über- oder Unterschätzung des Studienerfolgs hin. Ebenso zeigt sich, dass ältere und jüngere Studierfähigkeitstestversionen gleichermaßen von der geschlechtsspezifischen differenziellen Prognose betroffen sind. Dieser Befund wiederlegt die Hypothese, ältere Testverfahren würden sich häufiger mit typisch männlichen Themen befassen, wodurch sie das Potential von Frauen unterschätzen (Zwick, 2002, S. 151). Mittelwertsunterschiede im Studienerfolg sind ebenfalls kein Moderator für das Auftreten von differenzieller Prognose, was gegen die Annahme spricht, dass Frauen unterschätzt werden, weil sie generell bessere Noten erhalten als ihre männlichen Kommilitonen.

Einer der populärsten Erklärungsansätze für die Unfairnessbefunde ist das unterschiedliche Kurswahlverhalten von Männern und Frauen (Alon & Gelbgiser, 2011; Berry & Sackett, 2009). Die eingesetzten deutschen Studierfähigkeitstests unterschätzen jedoch die Studienleistungen von Frauen, auch wenn keine Kurswahlmöglichkeiten innerhalb der Studienfächer bestehen und damit die Notenzusammensetzung konstant gehalten wird (vgl. Studie 3). Demnach ist das Kurswahlverhalten, zumindest in Deutschland, nicht (allein) verantwortlich für die Unterschätzung der Studienleistungen von Frauen. Ähnlich fällt das Fazit für *stereotype threat*-Einflüsse während der Testsituation aus: auch wenn *stereotype threat* bei der Testdurchführung vermieden wird, zeigt sich eine Unterschätzung der Studienleistungen von Frauen (vgl. Studie 3).

Analog zu den Befunden von Kuncel und Hezlett (2007) ergibt sich für *Undergraduate-Tests* eine stärkere Unterschätzung der Frauen als für *Graduate-Tests* (Studie 1), obwohl beide Testarten sich in ihrem inhaltlichen Aufbau kaum

unterscheiden (vgl. Abschnitt 2.2). Vielmehr scheinen Frauen und Männer im *Graduate-Studium* sich in ihren Persönlichkeitseigenschaften weniger zu unterscheiden als Frauen und Männer im *Undergraduate-Studium*. Diese Annahme erhält Unterstützung von den Ergebnissen aus Studie 3. Für naturwissenschaftliche Studienanfänger verringert sich die differenzielle Prognose, wenn die Persönlichkeitsmerkmale Selbstdisziplin oder Leistungsmotivation bei der Vorhersage berücksichtigt werden.

Zusammenfassend macht die vorliegende Arbeit deutlich, dass weder das Testalter, noch Mittelwertsunterschiede im Test oder im Kriterium, das Kurswahlverhalten oder *stereotype threat* die Unterschätzung der Studienleistungen von Frauen durch Studierfähigkeitstests erklären können. Am vielversprechendsten scheinen Persönlichkeitsmerkmale wie Leistungsmotivation und Gewissenhaftigkeit, jedoch deuten die Ergebnisse auch darauf hin, dass nicht eine Erklärungsvariable allein für die geschlechtsspezifische differenzielle Prognose verantwortlich ist. Weitere aussichtsvolle Erklärungsansätze, die künftig näher untersucht werden sollten, werden am Ende des folgenden Abschnitts vorgestellt.

6.2 Zukünftige Forschung

Die durchgeführten Studien liefern vielfältige Anknüpfungspunkte für zukünftige Forschungsprojekte, mit dem Ziel weiter zu spezifizieren, unter welchen Bedingungen und mit welchem Ausmaß Frauen durch Studierfähigkeitstests unterschätzt werden und welche Einflussfaktoren diesen Effekt aufklären können.

Zunächst ist festzustellen, dass kaum empirische Arbeiten zur geschlechtsspezifischen differenziellen Prognose von *nicht*-amerikanischen Studierfähigkeitstests existieren und dass die bestehenden Studien oft methodische Einschränkungen aufweisen (vgl. Studie 1). Diese Lücke sollte geschlossen werden, indem Studierfähigkeitstests weltweit verstärkt auf ihre geschlechtsspezifische

differenzielle Prognose hin evaluiert werden. Mit Blick auf zukünftige Metaanalysen sollten Primärstudien sowohl geschlechtsspezifische Residuen als auch den exakten Verlauf der Regressionsgeraden für Männer und Frauen berichten.

Die starke Unterschätzung der Studienleistungen von Frauen im oberen Leistungsbereich sollte anhand großer Stichproben repliziert werden um zu prüfen, ob die gewonnen Erkenntnisse dieser Arbeit stabil sind und sich auch für andere Studierfähigkeitstests und andere Studienfächer zeigen. Neue Erkenntnisse aus den USA beschreiben unterschiedliche Ergebnisse für verschiedene Studienfächer. Der Studienerfolg von Frauen in technischen Fächern wird durch den SAT tendenziell unterschätzt, während der Test die Leistungen von Frauen in pädagogischen Fächern überschätzt (Shaw, Kobrin, Patterson & Mattern, 2011). Groß angelegte Studien könnten zeigen, inwiefern dieser Befund systematisch mit dem Auftreten von geschlechtsspezifischer differenzieller Prognose zusammenhängt. Darüber hinaus ist bisher wenig bekannt über den Einfluss der Applikationsmethode (Selbsttest vs. kontrollierte Testdurchführung) und des Freiwilligkeitsgrads (freiwillige Teilnahme vs. Test als Zulassungsbedingung). Diese Faktoren sollten in zukünftigen Forschungsarbeiten genauer betrachtet werden.

Bestehende Arbeiten haben bis jetzt darauf verzichtet, die zweite Hälfte des Studienabschnitts als Indikator für den Studienerfolg heranzuziehen, obwohl diese häufig mit anderen Erfolgskriterien verbunden ist als die ersten beiden Semester (Stemler, 2012). Um zu prüfen, ob die Genderunfairness für die Vorhersage des zweiten Studienabschnitts Bestand hat, sind besonders längsschnittliche Studien notwendig, die den Verlauf der differenziellen Prognose in Abhängigkeit vom Studienfortschritt aufzeigen.

Über die Frage, warum Frauen bessere Studienleistungen erzielen als es Studierfähigkeitstests vorhersagen, wurde in der Vergangenheit viel spekuliert. Viele der postulierten Hypothesen konnten mit der vorliegenden Arbeit widerlegt werden, weitere gilt es in folgenden Forschungsarbeiten zu berücksichtigen. Die Abiturnote ist hierfür

ein nützlicher Ausgangspunkt, da sie offenbar etwas misst, was den Studienerfolg von Frauen eher überschätzt und damit in Verbindung mit den Leistungen im Studierfähigkeitstest zu einer genauen Prognose führt (vgl. Studie 3). Gegebenenfalls ist der Abiturerfolg ein guter simulationsorientierter Indikator für das Lernverhalten während dem Studium. Tatsächlich finden sich erhebliche Unterschiede im Studierverhalten von Männern und Frauen. Frauen verbringen mehr Zeit mit Lernen, verfügen über ein besseres Zeitmanagement und nehmen öfters Unterstützung an als Männer (Marrs & Sigler, 2012; Wilson, 2007). Ebenso könnte studienfeld-spezifisches Vorwissen (wie es im Abitur abgefragt wird) eine entscheidende Rolle spielen. Die Ergebnisse von Mattern und Wyatt (2012) wiedersprechen doch dieser Erklärung. In ihrer Studie verringert die Berücksichtigung des Vorwissens nicht die geschlechtsspezifische differenzielle Prognose von Studierfähigkeitstests.

Interessant wäre die Entwicklung eines multidimensionalen Modells, welches den Zusammenhang von verschiedenen Persönlichkeitseigenschaften (wie Leistungsmotivation und Gewissenhaftigkeit), dem Studienverhalten, dem Geschlecht und dem Studienerfolg darstellt. Solch ein Modell könnte wichtige Hinweise für eine genderfaire Auswahl von Studierenden liefern. Zugleich stellt die Erklärung kleiner Effekte mit multifaktoriellen Modellen jedoch eine besondere Herausforderung dar.

Weitere Anknüpfungspunkte um zu erklären warum Frauen bessere Studienleistungen erzielen als es Studierfähigkeitstests vorhersagen sind die Testsituation sowie der Inhalt und die Aufgabenformate von Studierfähigkeitstests. Steele und Aronson (1995) belegen, dass alleine die Test-Instruktion zur Aktivierung von *stereotype threat* bei Farbigen führen kann. Wird in der Instruktion von einem Leistungstest gesprochen, fällt das Ergebnis der Farbigen deutlich geringer aus als wenn kein Leistungsbezug hergestellt wird. Geschlechtsspezifische Instruktions-Effekte zeigen sich vor allem für Mathematikaufgaben. Frauen schneiden in Mathematiktests schlechter ab, wenn in der Instruktion ein Leistungsbezug hergestellt wird als wenn keine spezifische Instruktion erfolgt (Marx & Stapel, 2006). Experimentelle Studien

sollten überprüfen, ob dieser Effekt auch bei Studierfähigkeitstests zum Tragen kommt, wenn weibliche Versuchsleiterinnen eingesetzte werden.

Auch das Multiple-Choice-Aufgabenformat wird als Hürde für Frauen vermutet (Bolger & Kellaghan, 1990; Bridgeman & Lewis, 1994; Lindberg et al., 2010). Frauen scheinen vor allem bei Speed-Aufgaben schlechter abzuschneiden (Goldstein et al., 1990), da sie weniger Aufgaben bearbeiten als Männer. Bis heute ist unklar, ob ein gewissenhafteres Antwortverhalten oder der ansteigende Schwierigkeitsgrad der Items hierfür verantwortlich ist (Mäkitalo, 1996; Åberg-Bengtsson, 1999).

Nicht zuletzt wurden in der Vergangenheit wenige Bemühungen unternommen Aufgabenformate zu entwickeln, die beiden Geschlechtern gleichermaßen gerecht werden und es wurde kaum untersucht, ob vielleicht unterschiedliche Facetten des schlussfolgernden Denkens für den Studienerfolg von Männern und Frauen verantwortlich sind, zumal noch Spielraum in der Varianzausschöpfung besteht (Patterson et al., 2009; Patterson & Mattern, 2011). Eine bewusstere Auswahl der Aufgabenformate könnte die Genderfairness der Tests verbessern.

Aus einem allgemeineren Blickwinkel stellt sich die Frage, wie die differenzielle Prognose mit den anderen statistischen Fairnessmerkmalen genau zusammenhängt (vgl. Abschnitt 2.5). DIF, differenzielle Validitäten und differenzielle Prädiktion sind nicht gänzlich unabhängig voneinander, es fehlen jedoch Modelle und Simulationsstudien, welche die Verbindungen der einzelnen Fairnessaspekte aufzeigen und konkretisieren. Die Isolation von Items, welche geschlechtsspezifisches DIF zeigen, könnte sich günstig auf die geschlechtsspezifische differentielle Prognose auswirken (Päßler, Leidig & Hell, 2012), ein Ansatz der bei zukünftigen Testentwicklungen aufgegriffen werden sollte.

6.3 Implikationen für die Testentwicklung

Es wäre zu begrüßen, wenn sich in Zukunft bei Testentwicklern (und Anwendern) eine genderfaire differenzielle Prognose der Tests als ein von der Bedeutung her mit der Kriteriumsvalidität vergleichbares Qualitätsmerkmal etablieren würde. Dann könnte bereits während der Testkonstruktion das Problem der differenziellen Prognose unterbunden werden.

Wie im vorangegangenen Abschnitt erläutert, sollten neue Aufgabenformate erprobt werden, mit dem Ziel die Genderfairness und optimaler Weise auch die Validität zu verbessern. Weniger Speed-Komponenten und mehr offene Antwortformate könnten helfen einen genderfairen Test zu konstruieren. Ebenso sollte versucht werden mehr verbale Testteile in den Studierfähigkeitstest aufzunehmen, da verbale Testteile weniger differenzielle Prognose zeigen als mathematische Testteile (vgl. Studie 1), auch wenn sich für den SAT gezeigt hat, dass sich mehr verbale Teststeile nicht zwingend günstig auf die geschlechtsspezifische Fairness auswirken (Patterson et al., 2009).

Bei der Testkonstruktion sollte nicht eine bloße Nivellierung der geschlechtsspezifischen Mittelwertsunterschiede im Test angestrebt werden, da die Ergebnisse der Metaanalyse zeigen, dass dies nicht zwingend das Auftreten von geschlechtsspezifischer differenzieller Prognose verhindert (vgl. Studie 1). Vielmehr sollten einzelne Items bzw. Aufgabenformate daraufhin überprüft werden, ob sie den Studienerfolg für beide Geschlechter gleichermaßen fair vorhersagen können oder ob sie durch weitere Aufgaben zur Erfassung von Leistungsmotivation bzw. Gewissenhaftigkeit ergänzt werden sollten.

6.4 Implikationen für die Studienberatung und die Auswahl von Studierenden

Aus den gewonnen Ergebnissen dieser Arbeit lassen sich wichtige Empfehlungen für die Verwendung von Studierfähigkeitstests und Abiturnoten im Rahmen der Hochschulzulassung und Studienberatung ableiten. Der Einsatz von Studierfähigkeitstests kann dazu führen, dass Frauen unter Umständen von einem Studium abgeraten wird bzw. sie von den Universitäten keinen Studienplatz angeboten bekommen, obwohl sie eigentlich gute Studienleistungen erbringen würden. Ähnliches kann aber auch Männern passieren, wenn für die Beratung bzw. Auswahl von Studierenden ausschließlich auf die Abiturnoten zurückgegriffen wird. Hieraus ergeben sich unterschiedliche Implikationen für die Beratung und Auswahl von Studierenden, in Abhängigkeit von der Selektivität des betreffenden Studienfachs.

Für nicht zulassungsbeschränkte Studiengänge können Studierfähigkeitstests im Rahmen einer Studienberatung den Studieninteressierten eine objektive Rückmeldung darüber geben, wie ausgeprägt ihre für das Studienfeld relevanten kognitiven Fähigkeiten sind (Kubinger et al., 2012). Diese Information ist insbesondere hilfreich für Ratsuchende, deren Abiturleistung eher unterdurchschnittlich ausfällt, Quereinsteiger und Bewerber, die aus bestimmten Gründen während der Abiturphase nicht ihr eigentliches Leistungspotential abrufen konnten. Aus Sicht der Genderfairness ist es unproblematisch die Tests in dieser Weise einzusetzen, da sich für den mittleren bis unteren Leistungsbereich keine unfaire Vorhersage zeigt.

Ein anderes Bild ergibt sich für stark zulassungsbeschränkte Studienfächer bzw. Studieninteressierte mit überdurchschnittlichen Abiturleistungen. Hier geht es darum den Ratsuchenden eine Einschätzung zu geben, inwiefern sie mit anderen leistungsstarken Studienanfängern mithalten können. Für weibliche Ratsuchende sollte sich die Empfehlung hauptsächlich an den Abiturnoten orientieren und die Leistungen im Studierfähigkeitstest sollten nur (wenn überhaupt) mit einer geringen Gewichtung berücksichtigt werden, außer es bestanden leistungsmindernde Umstände während der

Abiturphase. Männliche Ratsuchende mit Interesse an stark zulassungsbeschränkten Studiengängen und durchschnittlichen Abiturnoten sollten hingegen ermutigt werden ihr Fähigkeitsvermögen zusätzlich durch einen Studierfähigkeitstest zu klären, gleichzeitig sollten sie jedoch darauf hingewiesen werden, dass ein Studium ähnliche Anforderungen an das Lernverhalten stellt wie die Schule.

Für die Hochschulzulassung in Deutschland sind Schulnoten als Einzelprädiktor das am häufigsten verwendete Auswahlkriterium (Heine, Briedis, Didi, Haase & Trost, 2006). Für streng selektive Studienfächer ist dieses Vorgehen kritisch zu bewerten, da es eine mögliche Unterschätzung der Männer in Kauf nimmt. Fächer mit strengen Zulassungsbeschränkungen sollten deshalb prüfen, ob zusätzlich zur Abiturnote Studierfähigkeitstests oder weitere Auswahlinstrumente berücksichtigt werden sollten um eine genderfaire Hochschulzulassung zu gewährleisten.

Die Auswahl von Studierenden anhand von Studierfähigkeitstests ist mit einem gewissen Aufwand und nicht zu vernachlässigenden Kosten verbunden. Gleichzeitig steigt die Validität und damit der Nutzen von Studierfähigkeitstests mit der Strenge der Selektionsquote (Hell et al., 2008). Aus diesen Gründen werden Studierfähigkeitstests gewöhnlich nur bei sehr strikten Auswahlentscheidungen eingesetzt. Dies ist zunächst aus Sicht der Genderfairness als ungünstig zu bewerten. In der Praxis werden Studierfähigkeitstests aus rechtlichen Gründen jedoch gewöhnlich in Verbindung mit der Abiturleistung als kombiniertes Zulassungskriterium verwendet. Dieses Vorgehen führt zu einer validen und gleichzeitig auch genderfairen Vorhersage und ist damit zu begrüßen.

Literaturverzeichnis

* Mit einem Stern versehene Referenzen bezeichnen Forschungsarbeiten, die in die Metaanalyse (Kapitel 3) aufgenommen wurden.

Åberg-Bengtsson, L. (1999). Dimensions of performance in the interpretation of diagrams, tables, and maps: Some gender differences in the Swedish Scholastic Aptitude Test. *Journal of Research in Science Teaching, 36*, 565-582.

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., 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

Aloe, A. M. & Becker, B. J. (2009). Teacher verbal ability and school outcomes: Where is the evidence? *Educational Researcher, 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

Amelang, M. & Funke, J. (2005). Entwicklung und Implementierung eines kombinierten Beratungs- und Auswahlverfahrens für die wichtigsten Studiengänge an der Universität Heidelberg. *Psychologische Rundschau, 56*, 135-137.

*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*. Washington, DC: American Educational Research Association.

Arnhold, N. & Hachmeister, C.-D. (2004). *Leitfaden für die Gestaltung von Auswahlverfahren an Hochschulen* (Arbeitspapier Nr. 52). Gütersloh: Centrum für Hochschulentwicklung.

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.

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

Berkowitz, D. & Hoekstra, M. (2011). Does high school quality matter? Evidence from admissions data. *Economics of Education Review, 30*, 280-288.

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

Blakemore, J. E. O., Berenbaum, S. A. & Liben, L. S. (2009). *Gender development*. New York, NY: Psychology Press.

Blum, F. (1997). Zahlenmäßige Anteile, Test- und Schulleistungen einzelner Gruppen von Testteilnehmern. In G. Trost (Hrsg.), *Test für Medizinische Studiengänge (TMS): Studien zur Evaluation (21. Arbeitsbericht)* (S. 34-74). Bonn: ITB.

Bolger, N. & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. *Journal of Educational Measurement, 27*, 165-174. doi: 10.1111/j.1745-3984.1990.tb00740.x

Borenstein, M., Hedges, L. V., Higgins, J. P. T. & 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

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). New York, NY: The College Board.

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

*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

*Burton, N. W. & Wang, M. (2005). *Predicting long-term success in graduate school: A collaborative validity study* (GRE Board Research Report No. 99-14R). Princeton, NJ: Educational Testing Service.

*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)

Calvin, C. M., Fernandes, C., Smith, P., Visscher, P. M. & Deary, I. J. (2010). Sex, intelligence and educational achievement in a national cohort of over 175,000 11-year-old schoolchildren in England. *Intelligence, 38*, 424-432.

*Casserly, P. L. (1982). *Older students and the SAT* (College Board Report No. 82-2). New York, NY: The College Board.

Chamorro-Premuzic, T., Harlaar, N., Greven, C. U. & Plomin, R. (2010). More than just IQ: A longitudinal examination of self-perceived abilities as predictors of academic performance in a large sample of UK twins. *Intelligence, 38*, 385-392.

Chemers, M. M., Hu, L. & Garcia, B. F. (2001). Academic self-efficacy and first year college student performance and adjustment. *Journal of Educational Psychology, 93,* 55-64.

*Chou, T. & Huberty, C. J. (1990). *A freshman admissions prediction equation: An evaluation and recommendation.* Retrieved from ERIC database. (ED333081)

*Clark, M. J. & Grandy, J. (1984). *Sex differences in the academic performance of Scholastic Aptitude Test takers* (College Board Report No. 84-8). New York, NY: The College Board.

Cleary, T. A. (1968). Test bias: 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.

College Board. (2011). *2011 College-bound seniors: Total group profile report*. New York, NY: Author.

Colom, R. & Lynn, R. (2004). Testing the developmental theory of sex differences in intelligence on 12-18 year olds. *Personality and Individual Differences, 36*, 75-82.

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). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, *4*, 5-13.

Costa, P. T. & McCrae, R. R. (1998). Six approaches to the explication of facet-level traits: Examples from conscientiousness. *European Journal of Personality, 12*, 117-134.

*Cowen, S. & Fiori, S. J. (1991, November). *Appropriateness of the SAT in selecting students for admission to California State University, Hayward*. Paper presented at the annual meeting of the California Educational Research Association, San Diego, CA. Retrieved from ERIC database. (ED343934)

*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. (ED 326541)

Cronbach, L. J. (1949). *Essentials of psychological testing*. Oxford, England: Harper.

Curley, E. W. & Schmitt, A. P. (1993). *Revising SAT-verbal items to eliminate differential item functioning* (College Board Report No. 93-2). New York, NY: The College Board.

Day, L., Hanson, K., Maltby, J., Proctor, C. & Wood, A. (2010). Hope uniquely predicts objective academic achievement above intelligence, personality, and previous academic achievement. *Journal of Research in Personality, 44*, 550-553.

Deary, I. J., Strand, S., Smith, P. & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*, 13-21.

De Fruyt, F., Van Leeuwen, K., De Bolle, M. & De Clercq, B. (2008). Sex differences in school performance as a function of conscientiousness, imagination and the mediating role of problem behaviour. *European Journal of Personality, 22*, 167-184.

Deidesheimer Kreis. (1997). *Hochschulzulassung und Studieneingangstests: Studienfeldbezogene Verfahren zur Feststellung der Eignung von Numerus-clausus- und andere Studiengänge*. Göttingen: Vadenhoeck & Ruprecht.

De Raad, B. & Schouwenburg, H. C. (1996). Personality in learning and education: A review. *European Journal of Personality, 10*, 303-336.

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

Downey, D. B. & Vogt Yuan, A. S. (2005). Sex differences in school performance during high school: Puzzling patterns and possible explanations. *Sociological Quarterly, 46*, 299-321.

Duckworth, A. L. & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science, 16*, 939-944.

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.

Eccles, J. S. & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*(1), 109-132.

Educational Testing Service. (2011). *GRE 2011-2012: Guide to the use of scores*. Princeton, NJ: Author.

*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 et al. (2008). *Sex differences: Summarizing more than a century of scientific research*. New York, NY: Psychology Press.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin, 116*, 429-456.

Fischer, F., Schult, J. & Hell, B. (2012a). Sex-specific differential prediction of college admission tests: A meta-analysis. Manuscript submitted for publication.

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

Fletcher, J. & Tienda, M. (2010). Race and ethnic differences in college achievement: Does high school attended matter? *Annals of the American Academy of Political and Social Science, 627*, 144-166.

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

Freudenthaler, H. H., Spinath, B. & Neubauer, A. C. (2008). Predicting school achievement in boys and girls. *European Journal of Personality, 22*, 231-245.

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.

Furnham, A. & Monsen, J. (2009). Personality traits and intelligence predict academic school grades. *Learning and Individual Differences, 19*, 28-33.

Furnham, A., Monsen, J. & Ahmetoglu, G. (2009). Typical intellectual engagement: Big five personality traits, approaches to learning and cognitive ability predictors of academic performance. *British Journal of Educational Psychology, 79*, 769-782.

Gavin, M. K. & Reis, S. M. (2005). Helping teachers to encourage talented girls in mathematics. In S. K. Johnsen & J. Kendrick (Eds.), *Teaching and counseling gifted girls.* (pp. 147-167). Waco, TX: Prufrock Press.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. et al. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*, 84-96.

Goldstein, D., Haldane, D. & Mitchell, C. (1990). Sex differences in visual-spatial ability: The role of performance factors. *Memory & Cognition, 18*, 546-550. doi: 10.3758/BF03198487

Gottfredson, L. S. (2003). G, jobs and life. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 293-342). Oxford, England: Elsevier.

Graduate Management Admission Council (2012). *World geographic trend report for GMAT Examinees*. Reston, VA: Author.

Grand, J. A., Ryan, A. M., Schmitt, N. & Hmurovic, J. (2011). How far does stereotype threat reach? The potential detriment of face validity in cognitive ability testing. *Human Performance, 24*, 1-28. doi: 10.1080/08959285.2010.518184

Grigorenko, E. L., Jarvin, L., Diffley, R., Goodyear, J., Shanahan, E. J. & Sternberg, R. J. (2009). Are SSATS and GPA enough? A theory-based approach to predicting academic success in secondary school. *Journal of Educational Psychology, 101*, 964-981.

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*, 91-114. doi: 10.1007/BF02289195

Haase, K. (2008). Studierendenauswahl im internationalen Vergleich. In H. Schuler & B. Hell (Hrsg.), *Studierendenauswahl und Studienentscheidung* (S. 43-54). Göttingen: Hogrefe.

Halpern, D. F. (2000). *Sex differences in cognitive abilities* (3rd ed.). Mahwah, NJ: Erlbaum.

Hannover, B. & Kessels, U. (2011). Sind Jungen die neuen Bildungsverlierer? Empirische Evidenz für Geschlechterdisparitäten zuungunsten von Jungen und Erklärungsansätze. *Zeitschrift für Pädagogische Psychologie, 25*, 89-103.

Hänsgen, K.-D., Spicher, B., Mallinger, R., Holzbaur, C., Dietrich, M. & Heidegger, M. (2008). *EMS Eignungstest für das Medizinstudium in Österreich 2008* (Bericht des Zentrums für Testentwicklung und der Medizinischen Universitäten Wien und Innsbruck). Fribourg: Zentrum für Testentwicklung.

Heckman, J. J. & LaFontaine, P. A. (2007). *The American high school graduation rate: Trends and levels.* (Discussion paper No. 3216). Bonn: IZA. http://ftp.iza.org/dp3216.pdf. Accessed 15 September 2011.

Hedges, L. V. & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*, 41-45.

Heine, C., Briedis, K., Didi, H.-J., Haase, C. & Trost, G. (2006). *Bestandsaufnahme von Auswahl- und Eignungsfeststellungsverfahren beim Hochschulzugang in Deutschland und ausgewählten Ländern* (HIS-Kurzinformation A 3/2006). Hannover: HIS.

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*(3), 251-270.

Hell, B., Trapmann, S. & Schuler, H. (2008). Synopse der Hohenheimer Metaanalysen zur Prognostizierbarkeit des Studienerfolgs und Implikationen für die Auswahl- und Beratungspraxis. In H. Schuler & B. Hell (Hrsg.), *Studierendenauswahl und Studienentscheidung* (S. 43-54). Göttingen: Hogrefe.

Hell, B., Trapmann, S., Weigand, S. & Schuler, H. (2007). Die Validität von Auswahlgesprächen im Rahmen der Hochschulzulassung - eine Metaanalyse. *Psychologische Rundschau, 58*, 93-102.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*, 47-77.

Heublein, U., Hutzsch, C., Schreiber, J., Sommer, D. & Besuch, G. (2009). *Ursachen des Studienabbruchs in Bachelor- und in herkömmlichen Studiengängen: Ergebnisse einer bundesweiten Befragung von Exmatrikulierten des Studienjahres 2007/08*. Hannover: HIS.

*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, A., Schumacher, J., Albani, C., Schmid, G. & Brähler, E. (2006). Bevölkerungsrepräsentative Normierung der Skala zur Allgemeinen Selbstwirksamkeitserwartung. *Diagnostica, 52*, 26-32.

Hodapp, V. (1991). Das Prüfungsängstlichkeitsinventar TAI-G: Eine erweiterte und modifizierte Version mit vier Komponenten. *Zeitschrift für Pädagogische Psychologie, 5*, 121-130.

Hodapp, V., Rohrmann, S. & Ringeisen, T. (2011). *Prüfungsangstfragebogen (PAF)*. Göttingen: Hogrefe.

*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

*House, J. D. (1998, May). *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)

*House, J. D. & Keeley, E. J. (1993, November). *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)

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.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist, 60*, 581-592.

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.

ITB Consulting GmbH (2012). *Test für medizinische Studiengänge* [Informationsbroschüre]. Bonn: Autor.

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.

Jäger, R. (2005). Zur Auswahl von Studierenden - einige Gedanken und Bedenken. *Psychologische Rundschau, 56*, 144-146.

Jensen, A. R. (1980). *Bias in mental testing*. New York, NY: Free Press.

*Jones, R. F. & Vanyur, S. (1985, April). *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)

*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., 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). New York, NY: The College Board.

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

Konegen-Grenier, C. (2001). *Studierfähigkeit und Hochschulzugang* (Band 61). Köln: Deutscher Instituts-Verlag.

Kubinger, K. D., Frebort, M. & Müller, C. (2012). Self-Assessment im Rahmen der Studienberatung: Möglichkeiten und Grenzen. In K. D. Kubinger, F. Martina, L. Khorramdel & L. Weitensfelder (Hrsg.), *Self-Assessment: Theorie und Konzepte* (S. 9-24). Lengerich: Pabst.

Kubinger, K. D., Moosbrugger, H., Frebort, M., Jonkisz, E. & Reiß, S. (2007). Die Bedeutung von Self-Assessments für die Studienplatzbewerbung. *Report Psychologie, 32,* 322-332.

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.

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: 10.1037/0033-2909.127.1.162

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

Kuncel, N. R., 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://etd.ohiolink.edu/. Accessed 12 September 2011.

Laidra, K., Pullmann, H. & Allik, J. (2007). Personality and intelligence as predictors of academic achievement: A cross-sectional study from elementary to secondary school. *Personality and Individual Differences, 42*, 441-451.

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

Leeson, P., Ciarrochi, J. & Heaven, P. C. L. (2008). Cognitive ability, personality, and academic performance in adolescence. *Personality and Individual Differences, 45*, 630-635.

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.

Lewis, J. C. & Hoover, H. D. (1987). Differential prediction of academic achievement in elementary and junior high school by sex. *The Journal of Early Adolescence, 7*, 107-115.

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

Light, R. J. & Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press. doi: 10.2307/1175260

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

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

*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)

Lu, L., Weber, H. S., Spinath, F. M. & Shi, J. (2011). Predicting school achievement from cognitive and non-cognitive variables in a Chinese sample of elementary school children. *Intelligence, 39*, 130-140.

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.

Maichle, U. & Meyer, M. (1997). Elfter sowie zwölfter und letzter Einsatz des TMS im besonderen Auswahlverfahren: Testteilnehmer, Testablauf, Testcharakteristika. In G. Trost (Hrsg.), *Test für medizinische Studiengänge (TMS): Studien zur Evaluation (21. Arbeitsbericht)* (S. 76-134). Bonn: ITB.

Mäkitalo, Å. (1996). Gender differences in performance on the DTM subtest in the Swedish Scholastic Aptitude Test as a function of item position and cognitive demands. *Scandinavian Journal of Educational Research, 40*, 189-201. doi: 10.1080/0031383960400301

Marrs, H. & Sigler, E. A. (2012). Male academic performance in college: The possible role of study strategies. *Psychology of Men & Masculinity*, *13*, 227-241. doi: 10.1037/a0022247

Marx, D. M. & Stapel, D. A. (2006). Distinguishing stereotype threat from priming effects: On the role of the social self and threat-based concerns. *Journal of Personality and Social Psychology*, *91*, 243-254. doi: 10.1037/0022-3514.91.2.243

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). New York, NY: The College Board.

Mattern, K. D. & Wyatt, J. N. (2012). *The Validity of the Academic Rigor Index (ARI) for Predicting FYGPA* (Research Report No. 2012-5). New York, NY: The College Board.

McCarthy, J. M. & Goffin, R. D. (2005). Selection test anxiety: Exploring tension and fear of failure across the sexes in simulated selection scenarios. *International Journal of Selection and Assessment, 13*, 282-295.

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

Moosbrugger, H., Jonkisz, E. & Fucks, S. (2006). Studierendenauswahl durch die Hochschulen: Ansätze zur Prognostizierbarkeit des Studienerfolgs. *Report Psychologie, 31* (3), 114-123.

Mouw, J. T. & Khanna, R. K. (1993). Prediction of academic success: A review of the literature and some recommendations. *College Student Journal, 27*(3), 328-336.

Multon, K. D., Brown, S. D. & Lent, R. W. (1991). Relation of self-efficacy beliefs to academic outcomes: A meta-analytic investigation. *Journal of Counseling Psychology, 38*, 30-38.

Murphy, R. J. (1982). Sex differences in objective test performance. *British Journal of Educational Psychology, 52*, 213-219.

*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)* (S. 76-134). Bonn: ITB.

Nguyen, H. H. & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93*, 1314-1334. doi: 10.1037/a0012702

Noftle, E. E. & Robins, R. W. (2007). Personality predictors of academic outcomes: Big five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology, 93*, 116-130.

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

*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)

Päßler, K., Beinicke, A. & Hell, B. (2012). *Gender related Differential Validity and Differential Prediction in Interest Inventories*. Manuscript submitted for publication.

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). New York, NY: The College Board.

*Patterson, B. F., Mattern, K. D. & Kobrin, J. L. (2009). *Validity of the SAT for predicting FYGPA: 2007 SAT validity sample* (Statistical Report No. 2009-1). New York, NY: The College Board.

*Patton, T. K. (1998). *Differential prediction of college performance between gender.* Retrieved from ERIC database. (ED029407)

*Pennock-Román, M. (1994). *College major and gender differences in the prediction of college grades* (Report No. 94-2). New York, NY: The College Board.

Petrides, K. V., Chamorro-Premuzic, T., Frederickson, N. & Furnham, A. (2005). Explaining individual differences in scholastic behaviour and achievement. *British Journal of Educational Psychology, 75*, 239-255.

Pike, G. R. & Saupe, J. L. (2002). Does high school matter? An analysis of three methods of predicting first-year grades. *Research in Higher Education, 43*, 187-207.

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

Psychologischer Dienst. (2011). *Studienfeldbezogene Beratungstests (SFBT): Probieren geht vor Studieren* [Broschüre]. Nürnberg: Bundesagentur für Arbeit.

*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., Lewis, C. & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (Report No. 93-1). New York, NY: The College Board.

*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.

Rindermann, H. (2005). Für ein bundesweites Auswahlverfahren von Studienanfängern über Fähigkeitsmessung. *Psychologische Rundschau, 56*, 127-129.

Rindermann, H. & Oubaid, V. (1999). Auswahl von Studienanfängern durch Universitäten - Kriterien, Verfahren und Prognostizierbarkeit des Studienerfolgs. *Zeitschrift für Differentielle und Diagnostische Psychologie, 20*, 172-191.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R. & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*, 261-288.

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

Sanber, S. R. & Millman, J. (1987, April). *Gender and race effects on standardized tests predictive validity: A meta-analytical study*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC. Retrieved from ERIC database. (ED286914)

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-Atzert, L. & Krumm, S. (2006). Professionelle Studierendenauswahl durch die Hochschulen: Wege und Irrwege. *Report Psychologie, 31*(6), 297-309.

Schmitt, A. P. & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement*, *27*, 67-81.

Schuler, H. & Prochaska, M. (2000). Entwicklung und Konstruktvalidierung eines berufsbezogenen Leistungsmotivationstests. *Diagnostica, 46*, 61-72.

Schuler, H., Prochaska, M. & Frintrup, A. (2001). *Leistungsmotivationsinventar (LMI)*. Göttingen: Hogrefe.

Schuler, H., Thornton, G. C., Frintrup, A. & Mueller-Hanson, R. (2004). *Achievement Motivation Inventory (AMI)*. Ashland, OH: Hogrefe & Huber.

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Göttingen: Hogrefe & Huber.

Schwarzer, R. & Jerusalem, M. (Eds.). (1999). *Skalen zur Erfassung von Lehrer- und Schülermerkmalen: Dokumentation der psychometrischen Verfahren im Rahmen der Wissenschaftlichen Begleitung des Modellversuchs Selbstwirksame Schulen*. Berlin: Freie Universität Berlin.

Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety Research, 4*(1), 27-41.

Shadish, W. R. & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper and L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261-281). New York: Russell Sage Foundation.

Shaw, E. J., Kobrin, J. L., Patterson, B. F. & Mattern, K. D. (2011). *The validity of the SAT for predicting cumulative grade point average by college major.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Shen, C. & Pedulla, J. J. (2000). The relationship between students' achievement and their self-perception of competence and rigour of mathematics and science: A cross-national analysis. *Assessment in Education: Principles, Policy & Practice, 7*, 237-253.

*Siegert, K. O. (2007). *Predicting success in graduate management doctoral programs* (GMAC Research Reports No. RR-07-10). McLean, VA: Graduate Management Admission Council.

*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

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

Sparfeldt, J. R., Buch, S. R. & Rost, D. H. (2010). Klassenprimus bei durchschnittlicher Intelligenz: Overachiever auf dem Gymnasium. *Zeitschrift für Pädagogische Psychologie, 24*, 147-155.

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

Spiel, C., Schober, B. & Litzenberger, M. (2008). *Evaluation der Eignungstests für das Medizinstudium in Österreich.* Projektbericht für das Bundesministerium für Wissenschaft und Forschung, Wien.

Spinath, B., Freudenthaler, H. H. & Neubauer, A. C. (2010). Domain-specific school achievement in boys and girls as predicted by intelligence, personality and motivation. *Personality and Individual Differences, 48*, 481-486.

Spinath, B., Spinath, F. M., Harlaar, N. & Plomin, R. (2006). Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value. *Intelligence, 34*, 363-374.

Spinath, F. M., Spinath, B. & Plomin, R. (2008). The nature and nurture of intelligence and motivation in the origins of sex differences in elementary school achievement. *European Journal of Personality, 22*, 211-229.

Statistisches Bundesamt. (2010). *Bildung und Kultur: Allgemeinbildende Schulen.* (Fachserie 11 Reihe 1 Nr. 2110100107004). Wiesbaden: Autor.

Statistisches Bundesamt. (2011). *Bildung und Kultur: Erfolgsquoten 2009 Berechnung für die Studienanfängerjahrgänge 1007 bis 2001*. Wiesbaden: Autor.

Statistisches Bundesamt. (2012). *Bildung und Kultur: Studierende an Hochschulen* (Fachserie 11 Reihe 4.1 Nr. 2110410128004). Wiesbaden: Autor.

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

Steele, C. M. & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*, 797-811. doi: 10.1037/0022-3514.69.5.797

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.

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

Steinmayr, R. & Spinath, B. (2008). Sex differences in school achievement: What are the roles of personality and achievement motivation? *European Journal of Personality, 22*, 185-209.

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

Stemler, S. E. (2012). What should university admissions tests predict? *Educational Psychologist*, *47*, 5-17.

Sternberg, R. J., Grigorenko, E. L. & Bundy, D. A. (2001). The predictive value of IQ. *Merrill-Palmer Quarterly, 47*(1), 1-41.

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.

Strand, S., Deary, I. J. & Smith, P. (2006). Sex differences in cognitive abilities test scores: A UK national picture. *British Journal of Educational Psychology, 76*, 463-480.

*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

*Swinton, S. S. (1987). *The predictive validity of the restructured GRE with particular attention to older students* (GRE Board Report No. 83-25P). Princeton, NJ: Educational Testing Service.

*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). McLean, VA: Graduate Management Admission Council.

*Thomas, C. L. (1973, February). *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)

*Thomas, C. L. (1979). Relative effectiveness of high school grades for predicting college grades: Sex and ability level effects. *Journal of Negro Education, 48*(1), 6-13. doi: 10.2307/2294611

Thorndike, R. L. (1949). *Personnel selection: Test and measurement techniques*. New York, NY: Wiley.

Trapmann, S., Hell, B., Hirn, J. O. & Schuler, H. (2007). Meta-analysis of the relationship between the big five and academic success at university. *Zeitschrift für Psychologie, 215*, 132-151.

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.

Trost, G. (2003). *Deutsche und internationale Studierfähigkeitstests: Arten, Brauchbarkeit, Handhabung*. Bonn: DAAD.

Trost, G. (2005). Studierendenauswahl durch die Hochschulen: Welche Verfahren kommen prinzipiell in Betracht, welche nicht? *Psychologische Rundschau*, *56*, 138-140.

Trost, G., Klieme, E. & Nauels, H. (1997). Prognostische Validität des Tests für medizinische Studiengänge (TMS). In T. Herrmann (Hrsg.), *Hochschulentwicklung: Aufgaben und Chancen* (S. 57-87). Heidelberg: Asanger.

van Langen, A., Rekers-Mombarg, L. & Dekkers, H. (2006). Sex-related differences in the determinants and process of science and mathematics choice in pre-university education. *International Journal of Science Education, 28*, 71-94.

Veldman, D. J. (1968). Effects of sex, aptitudes, and attitudes on the academic achievement of college freshmen. *Journal of Educational Measurement, 5*, 245-249. doi: 10.1111/j.1745-3984.1968.tb00634.x

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.

Willingham, W. W., Pollack, J. M. & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, *39*, 1-37.

*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). Princeton, NJ: Educational Testing Service.

Wilson, R. (2007). The new gender divide. *Chronicle of Higher Education, 21,* 36-39.

Wissenschaftsrat. (2004). *Empfehlungen zur Reform des Hochschulzugangs*. Berlin: Autor.

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

Wottawa, H. & Amelang, M. (1980). Einige Probleme der "Testfairness" und ihre Implikationen für Hochschulzulassungsverfahren. *Diagnostica*, *26*, 199-221.

Wyatt, J., Wiley, A., Camara, W. J. & Proestler, N. (2012). *The development of an index of academic rigor for college readiness* (Research Report No. 2011-11). New York, NY: The College Board.

*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.* (Report No. 2001-6). New York, NY: The College Board.

*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

Zeidner, M. (1998). *Test anxiety: The state of the art*. New York, NY: Plenum Press.

Zimmerhofer, A., Heukamp, V. & Hornke, L. (2006). Ein Schritt zur fundierten Studienfachwahl: Webbasierte Self-Assessments in der Praxis. *Report Psychologie*, *31*(2), 62-72.

Zimmerhofer, A. & Trost, G. (2008). Auswahl- und Feststellungsverfahren in Deutschland - Vergangenheit, Gegenwart und Zukunft. In H. Schuler & B. Hell (Hrsg.), *Studierendenauswahl und Studienentscheidung* (S. 32-42). Göttingen: Hogrefe.

Zumdick, W. (2007). *Personality, sensation seeking and holiday preference*. Unpublished Bachelorthese, University of Twente, Enschede.

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

Zwick, R. & Himelfarb, I. (2011). The effect of high school socioeconomic status on the predictive validity of SAT scores and high school grade-point average. *Journal of Educational Measurement, 48*, 101-121.

Nachweis der Eigenleistungen

Die vorliegende Arbeit ist Teil meiner (FF) Arbeit als wissenschaftliche Mitarbeiterin und Doktorandin in dem Forschungsprojekt *Genderfairness berufs- und studieneignungsdiagnostischer Tests*, gefördert vom Bundesministerium für Bildung und Forschung sowie aus Mitteln des Europäischen Sozialfonds der Europäischen Union (FKZ01FP0930), unter der Leitung von Prof. Dr. Benedikt Hell (BH). Im Rahmen des Projekts bestand eine Zusammenarbeit mit Johannes Schult (JS), der ebenfalls als wissenschaftlicher Mitarbeiter und Doktorand angestellt war. Im Folgenden werden die einzelnen Beiträge von (BH), (JS) und (FF) genauer erläutert.

BH entwickelte die Ideen für die erste und die dritte Studie (Kapitel 3 und 5). FF und JS beteiligten sich an der genauen Konzeption des jeweiligen Studiendesigns. FF entwickelte die Idee für die zweite Studie (Kapitel 4). FF und JS waren, unter der Leitung von (BH), für die Durchführung der ersten, zweiten und dritten Studie (Kapitel 3-5) verantwortlich. Dies beinhaltete im Rahmen der Metaanalyse (Kapitel 3) die Entwicklung des Kodierschemas, die Literatursuche und die Kodierung der Studien. Im Rahmen der Längsschnittstudie (Kapitel 4 und 5) umfasste dies die Kooperation mit verschiedenen Fachbereichen, die Rekrutierung von Versuchspersonen und die Datenerhebung. Die statistische Auswertung der Daten (Kapitel 3-5) erfolgte durch FF, mit Einflüssen von JS und BH. JS war hauptsächlich für die Konzeption der statistischen Auswertungen der ersten Studie verantwortlich. Die Manuskripte von der ersten, zweiten und dritten Studie wurden von FF mit hilfreichem Input von JS und BH verfasst.