Luisa BERGHOLZ[1] & Stephan Josef STEGT (Bonn)

Validität und Fairness eines Studierfähigkeits-tests für Rechtswissenschaften

Zusammenfassung

Die vorliegende Studie untersucht ein schriftliches Auswahlverfahren für Rechtswissenschaften, bestehend aus der Abiturdurchschnittsnote und einem fachspezifischen Studierfähigkeitstest, im Hinblick auf seine prognostische Validität und seine Geschlechterfairness. Die Geschlechterfairness wird mit zwei verschiedenen Methoden geprüft. Die Daten von 579 Absolventinnen/Absolventen der Bucerius Law School werden analysiert. Die Ergebnisse zeigen, dass das Auswahlverfahren und der Auswahltest valide und fair sind. Auf Subtestebene zeigt sich ein differenziertes Bild: Die fachspezifischen und komplexen Subtests sind in dieser Studie valider und fairer als die unspezifischen Subtests.

Schlüsselwörter

Prognostische Validität, differentielle Validität, differentielle Prädiktion, Geschlechterfairness, Studierfähigkeitstest

[1] E-Mail: luisa.bergholz@itb-consulting.de

Wissenschaftlicher Beitrag · DOI: 10.3217/zfhe-13-04/04

Validity and fairness of an admissions test for law studies

Abstract

The present study examines the prognostic validity and gender fairness of a written admissions procedure for law studies in which the A-level is combined with a specific admissions test. Gender fairness is analyzed by means of two different methods. The data of 579 graduates of the Bucerius Law School are investigated. The results show that the admissions procedure and the admissions test are valid and fair. On the subtest level, the picture is more varied: the specific and complex subtests in this study are more valid and fair than the unspecific subtests.

Keywords

prognostic validity, differential validity, differential prediction, gender fairness, scholastic aptitude test

1 Hintergrund

Aufgrund hoher Bewerberzahlen setzen einige Hochschulen bei der Auswahl ihrer Studierenden mehrere Auswahlkriterien ein und kombinieren diese miteinander. Neben Schulnoten zählen fachspezifische Studierfähigkeitstests (SFTs) zu den validesten Prädiktoren des Studienerfolgs (HELL, TRAPMANN & SCHULER, 2007). Die höchste Validität wird häufig durch eine Kombination von beiden Prädiktoren erreicht (HELL, TRAPMANN & SCHULER, 2008). Fachspezifische SFTs finden außerdem hohe Akzeptanz bei Bewerberinnen/Bewerbern (z. B. HELL & SCHULER, 2005; HERDE, STEGT & PRECKEL, 2016).

Bei der Evaluation einiger internationaler Auswahltests wurde allerdings festgestellt, dass diese die Studienleistung von Frauen unterschätzen, beispielsweise der Scholastic Aptitude Test (SAT) (z. B. PATTERSON, MATTERN & KOBRIN, 2009; STRICKER, ROCK & BURTON, 1993; YOUNG & KOBRIN, 2001) sowie der Graduate Management Admission Test (GMAT) (HANCOCK, 1999). Eine

Zusammenfassung dieser Studien findet sich bei Fischer, Schult und Hell (2013). Hinweise, dass allgemeine kognitive Fähigkeitstests die Leistung von Frauen im Studium unterschätzen, lieferten Schult, Hell, Päßler und Schuler (2013).

Zur Frage, ob es eine Unterschätzung der Studienleistungen von Frauen auch bei fachspezifischen SFTs gibt, wie sie im deutschsprachigen Raum verwendet werden, liegen bislang wenige Untersuchungen vor, die noch kein klares Bild liefern. So zeigen Untersuchungen zum Test für medizinische Studiengänge (TMS) sowie zum Eignungstest für das Medizinstudium (EMS), dass die Tests keines der Geschlechter benachteiligen (NAUELS & MEYER, 1997; HÄNSGEN & SPICHER, 2001). Dlugosch (2005) beobachtete an einer kleinen Stichprobe für den Auswahltest der Bucerius Law School (BLS) hingegen eine leichte Unterschätzung der Leistung von Frauen. Fischer, Schult und Hell (2015) fanden in einer Studie mit nicht zulassungsrelevanten Beratungstests eine Unterschätzung der Leistung von Frauen im ersten Studienjahr, allerdings nur im oberen Leistungsbereich.

Bei der Konzeption der fachspezifischen SFTs im deutschsprachigen Raum wird Wert darauf gelegt, neben Aufgabentypen zur Messung allgemeiner kognitiver Fähigkeiten komplexere Aufgabentypen mit Fachbezug zu verwenden. Dieser Ansatz wird vom Deidesheimer Kreis (1997) folgendermaßen beschrieben:

„Durch die Auswahl der Aufgabentypen und die Ausgestaltung der Einzelaufgaben wird eine Simulation typischer Lern- und sonstiger Anforderungssituationen der betreffenden Studiengänge angestrebt. Soweit möglich, werden komplexe Aufgaben verwendet; auf die Einbeziehung homogener Aufgabengruppen, die jeweils nur eine einzige, eng umschriebene (Teil-)Fähigkeit messen, wird weitestgehend verzichtet. Komplexe Aufgaben werden dem Charakter kognitiver Leistungen im Studium, bei denen ganz verschiedene Fähigkeiten – teils kompensatorisch – eingesetzt werden müssen, am ehesten gerecht.“ (DEIDESHEIMER KREIS, 1997, S. 109-110)

Beim Einsatz solcher Aufgaben könnte die Gefahr, die Studienleistung einer Gruppe zu unter- oder überschätzen, geringer sein als bei Tests, die aus homogenen, nicht für die Komplexität der Studienanforderungen repräsentativen Aufgaben

bestehen. Denn wenn die Anforderungen im Test denjenigen im Studium ähneln, werden die für den Studienerfolg wichtigen Fähigkeiten in ihrem Zusammenspiel erfasst. Sofern in einem Test aber nur eine einzelne, eng umschriebene Fähigkeit erfasst wird, besteht die Möglichkeit, dass diese bei einer Personengruppe schwächer ausgeprägt ist. Die Personengruppe kann diese Schwäche im Studium unter Umständen mit nicht im Test erfassten Kompetenzen ausgleichen, so dass das reine Testergebnis die spätere Leistung unterschätzt.

Was die Autoren des vorliegenden Artikels mit komplexen, fachspezifischen Aufgabentypen im Gegensatz zu eindimensionalen und einfacher strukturierten Aufgabentypen meinen, sei anhand eines Beispiels und eines Gegenbeispiels veranschaulicht.

Beispiel: Beim Aufgabentyp „Fälle und Normen“ (siehe Abbildung 1) wird eine typische Anforderungssituation im Jurastudium simuliert: Es wird eine juristische Norm vorgegeben, die auf einen konkreten Einzelfall angewendet werden muss. Hier sind verbale Fähigkeiten und logisches Schlussfolgern im Rahmen einer komplexen und sorgfältigen Fallanalyse gefragt.

Gegenbeispiel: Beim Aufgabentyp „Zahlenreihen fortsetzen“ muss die testteilnehmende Person die Regel identifizieren, nach der eine Zahlenfolge aufgebaut ist, und die fehlende Zahl ergänzen, z. B. 1…2…4…7…11…16…___? In diesem Fall lautet die Regel +1, +2, +3, +4, +5 usw., die nächste Zahl ist die 22. Hier geht es um das Anwenden von Grundrechenarten und Erkennen von Rechenregeln. Dieser Aufgabentyp wird in einigen Intelligenztests verwendet (z. B. IST 2000 R, LIEPMANN, BEAUDUCEL, BROCKE & AMTHAUER, 2007).

Norm

§ 123 Strafgesetzbuch (StGB): Hausfriedensbruch
Hausfriedensbruch begeht, wer in die Wohnung, in die Geschäftsräume oder in abgeschlossene Räume, welche zum öffentlichen Dienst oder Verkehr bestimmt sind, widerrechtlich eindringt, oder wer, wenn er ohne Befugnis darin verweilt, auf die Aufforderung des Berechtigten sich nicht entfernt.

Sachverhalt

Die Firma SAFE kündigt zum Jahresende ordnungs- und fristgemäß ihren Vertrag mit der Firma Securitas, aufgrund dessen nächtliche Patrouillen am Bürogebäude von SAFE durch Securitas-Angestellte vorgenommen wurden. Im März des Folgejahres wird der Geschäftsführer von SAFE, Dr. Sefa, nachts um 3.00 Uhr von Securitas-Mitarbeitern angerufen: Ein Fenster im Bürogebäude war offen, die Securitas-Mitarbeiter, die durch ihren Chef nicht von der Kündigung durch SAFE unterrichtet worden waren, stiegen durch das Fenster ein und benachrichtigten Dr. Sefa und die Polizei. Dr. Sefa weist am Telefon darauf hin, dass der Vertrag mit Securitas seit langem gekündigt sei, und verlangt, dass die Securitas-Mitarbeiter sofort das Gebäude verlassen. Die Securitas-Mitarbeiter berufen sich indes auf ihren Auftrag und warten das Eintreffen der Polizei ab.

Welche der folgenden Aussagen lässt bzw. lassen sich aus einem Abgleich von Norm und Sachverhalt herleiten?

Aussage I: Die Securitas-Mitarbeiter haben Hausfriedensbruch begangen, als sie der telefonischen Aufforderung Dr. Sefas, das Bürogebäude zu verlassen, nicht Folge geleistet haben.

Aussage II: Der Chef der Securitas-Mitarbeiter hat Hausfriedensbruch begangen, da er billigend in Kauf genommen hat, dass seine Mitarbeiter widerrechtlich in Geschäftsräume eingedrungen sind.

(A) Nur Aussage I lässt sich ableiten.
(B) Nur Aussage II lässt sich ableiten.
(C) Beide Aussagen lassen sich ableiten.
(D) Keine der beiden Aussagen lässt sich ableiten.

Abb. 1: Beispielaufgabe „Fälle und Normen“

Mit der vorliegenden Studie sollen Validität und Geschlechterfairness eines zulassungsrelevanten Auswahltests mit Blick auf die Abschlussnoten im Studium untersucht werden. Dazu wird das schriftliche Auswahlverfahren der BLS überprüft. Entsprechend der Definition von Young und Kobrin (2001) werden differentielle Validität und differentielle Prädiktion wie folgt definiert: Differentielle Validität liegt vor, wenn die Korrelationen zwischen SFT und Kriterium für verschiedene Gruppen voneinander abweichen. Differentielle Prädiktion liegt vor, wenn die Regressionsgeraden und/oder die Standardschätzfehler für verschiedene Gruppen voneinander abweichen. Ein weiteres Ziel der Studie ist es, die verschiedenen Aufgabentypen des Auswahltests miteinander zu vergleichen und der Vermutung nachzugehen, dass mit fachspezifischen und komplexen Aufgabentypen im Sinne des Deidesheimer Kreises Validität und Fairness von Studierfähigkeitstests verbessert werden können.

2 Methode

2.1 Stichprobe

In den Jahren 2002 bis 2007 nahmen 2615 Bewerber/innen am Auswahlverfahren der BLS teil, davon 1126 Frauen (43 %). Von den Teilnehmenden, die infolge eines erfolgreichen Verfahrens ein Studium aufnahmen, schlossen 579 ihr Bachelorstudium (LL.B.) bis Anfang des Jahres 2012 erfolgreich ab, darunter 201 Frauen (35 %).

2.2 Auswahlverfahren und Auswahltest

Das Auswahlverfahren der BLS wurde im Jahr 2000 entwickelt. Der erste, schriftliche Teil des mehrstufigen Verfahrens besteht aus der Abiturdurchschnittsnote und dem Ergebnis im Auswahltest mit einer Gewichtung von 1:2. Jedes Jahr wird eine neue Version des Auswahltests eingesetzt. Die Teilnehmenden mit dem besten Gesamtergebnis im schriftlichen Auswahlverfahren werden zum mündlichen Teil

eingeladen. Knapp die Hälfte der Teilnehmenden am mündlichen Auswahlverfahren erhält eine Studienplatzzusage.

Der Test bestand aus den Subtests *Schlussfolgerungen, Diagramme und Tabellen, Indizien, Fälle und Normen* und *Oberbegriffe* (s. Tabelle 1). Bei *Schlussfolgerungen* müssen Behauptungen daraufhin geprüft werden, ob sie sich aus einer vorgegebenen Feststellung logisch ableiten lassen. Bei *Diagramme und Tabellen* gilt es, in Form von Grafiken und Tabellen dargestellte quantitative Informationen zu analysieren und zu interpretieren. Bei *Indizien* prüfen die Teilnehmenden, ob vorgegebene Thesen sich aus einem Satz von Informationen ableiten lassen bzw. mit diesem vereinbar sind. Bei *Fälle und Normen* sind Rechtsnormen auf konkrete Sachverhalte anzuwenden und bei *Oberbegriffe* sollen vorgegebene Wörter unter einen gemeinsamen Oberbegriff subsumiert und ein jeweils nicht passendes Wort erkannt werden. Für die Teilnehmenden gibt es eine Broschüre mit Beispielaufgaben.

Tab. 1: Aufbau des Tests.

Aufgabengruppe	Aufgaben-anzahl	Bearbeitungs-dauer
Schlussfolgerungen	22	21 Minuten
Diagramme und Tabellen	22	50 Minuten
Indizien	22	46 Minuten
Fälle und Normen	22	55 Minuten
Oberbegriffe	30	12 Minuten

Die Bearbeitungsdauer beträgt 184 Minuten; Teilnehmende können ein Testergebnis von 0 bis 94 Punkten erhalten. Die 579 LL.B.-Absolventinnen und Absolventen erreichten im Auswahltest im Durchschnitt 60.72 Punkte (SD = 8.35) und hatten eine durchschnittliche Abiturnote von 1.51 (SD = 0.38).

Die Aufgabengruppen *Schlussfolgerungen*, *Oberbegriffe* sowie *Diagramme und Tabellen* haben keinen direkten Bezug zum Studienfach. *Schlussfolgerungen* und *Oberbegriffe* liefern weitgehend homogene Messungen allgemeiner kognitiver Fähigkeiten (deduktives Schlussfolgern und verbale Fähigkeiten).

Die Aufgabengruppen *Indizien* und *Fälle und Normen* sind in hohem Maße fachspezifisch, weil sie eine in der Rechtswissenschaft grundlegende Denktechnik aufgreifen: die Subsumtion, d. h. die Anwendung einer Rechtsnorm auf einen Lebenssachverhalt. Rechtsnormen haben in der Regel die Wenn-Dann-Struktur eines klassischen Syllogismus, daher ist für ihre korrekte Anwendung die Verwendung aussagenlogischer Schlussregeln wichtig. In der Aufgabengruppe *Indizien* werden die Testteilnehmenden zunächst mit diesen Schlussregeln bekannt gemacht; dann müssen diese auf komplexe Wenn-Dann-Strukturen angewendet werden. Damit wird die der Subsumtion zugrundeliegende logische Denkform geprüft. Bei *Fälle und Normen* erfolgt die Subsumtion anhand echter Rechtsnormen und fiktiver Sachverhalte. Aufgrund der Komplexität der zu verarbeitenden Informationen prüfen die Aufgabengruppen darüber hinaus die für juristisches Denken notwendige Gewissenhaftigkeit und die Fähigkeit, viele Informationen im Gedächtnis zu halten und zu verknüpfen.

Diese beiden Aufgabengruppen entsprechen der Forderung des Deidesheimer Kreises (1997) nach fachspezifischen, komplexen Aufgaben in besonderem Maße. Das Testkonzept weist Ähnlichkeiten zu anderen fachspezifischen SFTs im deutschsprachigen Raum auf, die ebenfalls aus einer Kombination komplexer fachspezifischer und allgemeiner Aufgabengruppen bestehen.

2.3 Studienerfolgsmessung

Die Abschlussnote im Bachelor of Laws diente als Kriterium für den Studienerfolg. Die Notenskala reicht von 0 (ungenügend) bis 18 (sehr gut), zum Bestehen ist mindestens eine 4 (ausreichend) notwendig. Die Absolventinnen und Absolventen in dem untersuchten Zeitraum hatten eine durchschnittliche Abschlussnote von M = 9.56 (SD = 1.49).

2.4 Auswertungen

Die prognostischen Validitäten des Auswahltests, der Abiturnote sowie des gesamten schriftlichen Auswahlverfahrens wurden mit Produkt-Moment-Korrelationen berechnet. Zudem wurden schrittweise, vorwärts gerichtete multiple Regressionsanalysen durchgeführt, um die inkrementelle Validität der Prädiktoren zu bestimmen. Zur Ermittlung der differentiellen Validität wurden diese Berechnungen zunächst für die Gesamtgruppe und dann getrennt für Männer und Frauen durchgeführt.

Die Geschlechterfairness wurde wie folgt geprüft: Mit Fishers-z-Tests wurde zunächst ermittelt, ob sich die Korrelationskoeffizienten der beiden Gruppen unterscheiden. Anschließend wurde die differentielle Prädiktion mit dem Fairnessmodell von Cleary (1968) sowie dem Modell von Lawshe (1983) überprüft.

Laut Cleary (1968) ist ein Verfahren dann fair, wenn die Regressionsgeraden beider Gruppen deckungsgleich sind. Das heißt, die jeweilige Steigung, die Konstante und der Standardschätzfehler der Regressionsgeraden dürfen sich nicht signifikant voneinander unterscheiden. Unterscheiden sich die beiden Regressionsgeraden nur bezüglich der Konstante signifikant voneinander, dann wird die Gruppe mit dem größeren Achsenabschnitt in dem vorhergesagten Kriterium unterschätzt, die Gruppe mit dem kleineren Achsenabschnitt hingegen überschätzt. Bei dem Modell von Lawshe (1983) werden zunächst die Werte des Prädiktors und des Kriteriums für die Gesamtgruppe standardisiert und die Differenzen zwischen den jeweiligen standardisierten Prädiktor- und Kriteriumswerten berechnet. Anhand eines t-Tests wird dann geprüft, ob sich die Mittelwerte der Differenzen von den beiden unter-

suchten Gruppen signifikant unterscheiden. Ein Testverfahren gilt dann als unfair, wenn die Mittelwerte signifikant voneinander abweichen. Dabei steht ein positiver Wert für eine Überschätzung der Studienleistung, ein negativer Wert für eine Unterschätzung.

Alle Signifikanztests wurden mit einem Signifikanzniveau von $p = 0.05$ (zweiseitig) gerechnet. Zur Interpretation der Ergebnisse ist die Effektstärke d (Cohen, 1988) angegeben.

3 Ergebnisse

3.1 Deskriptive Statistiken

Tabelle 2 zeigt die deskriptiven Statistiken des Kriteriums und der Prädiktoren getrennt nach Geschlecht und das Ausmaß der geschlechtsspezifischen Unterschiede. Die Werte für Männer und Frauen unterscheiden sich signifikant voneinander hinsichtlich der Abiturnote, des Auswahltests sowie der Subtests *Schlussfolgerungen* und *Diagramme und Tabellen*. Die Frauen haben etwas bessere Abiturnoten, während die Männer etwas bessere Werte im Auswahltest insgesamt erzielen. Diese Geschlechtsunterschiede haben kleine bis mittlere Effektstärken. Bei den Subtests *Schlussfolgerungen* und *Diagramme und Tabellen* erreichten die Männer höhere Punktzahlen mit mittleren bis großen Effektstärken.

Tab. 2: Mittelwerte und Standardabweichungen des Kriteriums und der Prädiktoren getrennt nach Männern und Frauen, t-Werte, Effektstärke d.

	Männer		Frauen				
	M	SD	M	SD	t	p	d
1. LL.B.	9.57	1.49	9.54	1.48	0.19	.85	0.02
2. Abi	1.54	0.39	1.45	0.35	2.62	.01	0.24
3. Auswahltest	61.94	8.33	58.43	7.91	4.92	.00	0.43
4. Schlussflg	13.15	2.62	11.15	2.84	8.50	.00	0.73
5. DuT	11.44	2.72	9.78	2.58	7.09	.00	0.63
6. Indizien	11.07	3.54	11.57	3.46	-1.62	.11	-0.14
7. FuN	10.10	2.44	10.05	2.09	0.24	.81	0.02
8. Oberbegr	16.19	2.35	15.88	2.35	1.51	.13	0.13

Anmerkungen: $n_{Männer}$ = 378; n_{Frauen} = 201; LL.B. = Bachelor of Laws; Schlussflg = Subtest *Schlussfolgerungen*; DuT = Subtest *Diagramme und Tabellen*; Indizien = Subtest *Indizien*; FuN = Subtest *Fälle und Normen*; Oberbegr = Subtest *Oberbegriffe*. df (t-Test) = 577; Positive Effektstärken bedeuten eine höhere Merkmalsausprägung zu Gunsten der Männer; bei der Abiturnote verhält es sich umgekehrt, da es sich um eine invertierte Variable handelt.

3.2 Prognostische Validität

Die prognostische Validität gibt Auskunft darüber, wie gut das Auswahlverfahren den Studienerfolg vorhersagt. Dies lässt sich anhand von bivariaten Korrelationen zwischen den Prädiktoren und dem Kriterium überprüfen. Da es sich bei der Stichprobe nur um diejenigen Testteilnehmenden handelt, die zum Studium zugelassen wurden, müssen Varianzeinschränkungen auf Seiten des Prädiktors berücksichtigt werden. Diese Einschränkungen lassen sich mit Hilfe einer Selektionskorrektur ausgleichen. Hierzu wurde eine Korrekturformel von Thorndike (1949, zitiert nach LIENERT, 1967, S. 306) verwendet. Für diese Korrekturformel werden die Varianzen der Gesamtpopulation benötigt. Da den Autoren nur die Abiturnoten der ausgewählten Bewerber/innen bekannt sind und nicht diejenigen der Bewerber/innen, die zwar am Test teilgenommen, jedoch nicht zum Studium an der BLS zugelassen wurden, wurde die Varianz der Gesamtpopulation in diesem Fall anhand eines Verfahrens von Alexander, Alliger und Hanges (1984) geschätzt. Tabelle 3 zeigt die Korrelationen der Prädiktoren mit dem Kriterium mit und ohne Selektionskorrektur für die Gesamtstichprobe und getrennt nach Geschlecht.

Sowohl die Abiturnote als auch der Auswahltest und die einzelnen Subtests korrelieren signifikant mit der Studienleistung. Mit Selektionskorrektur kommt die Abiturnote alleine auf eine prognostische Validität von r_{tc} = .39 (15 % Varianzaufklärung), der Auswahltest erzielt eine prognostische Validität von r_{tc} = .46 (21 % Varianzaufklärung). Die Kombination aus Abiturdurchschnittsnote und Auswahltest hat mit Selektionskorrektur eine prognostische Validität von r_{tc} = .45 (20 % Varianzaufklärung) für das Kriterium LL.B.-Abschlussnote. Die Validitäten für Männer und Frauen unterscheiden sich nicht signifikant.

Tab. 3: Prognostische Validität der Abiturdurchschnittsnote und des Auswahltests für die Gesamtstichprobe sowie getrennt nach Geschlecht.

	Kriterium LL.B.					
Prädiktor	Gesamt unkorr.	Gesamt korr.	Männer unkorr.	Männer korr.	Frauen unkorr.	Frauen korr.
Abi	.39**	.39**	.41**	.41**	.37**	.37**
Auswahltest	.32**	.46**	.33**	.45**	.33**	.47**
Abi und Auswahltest (1:2-Gewichtung)	.45**	.45**	.45**	.47**	.46**	.49**
Schlussflg	.13**	.16**	.13**	.17**	.14	.15**
DuT	.19**	.23**	.19**	.23**	.20**	.22**
Indizien	.29**	.35**	.28**	.33**	.30**	.36**
FuN	.23**	.27**	.22**	.25**	.24**	.29**
Oberbegr	.12**	.15**	.13**	.17**	.08	.16**

Anmerkungen: n = 579; $n_{Männer}$ = 378; n_{Frauen} = 201; LL.B. = Bachelor of Laws; Schlussflg = *Schlussfolgerungen*; DuT = *Diagramme und Tabellen*; FuN = *Fälle und Normen*; Oberbegr = *Oberbegriffe*; unkorr. = unkorrigierte Korrelationskoeffizienten; korr. = mit Selektionskorrektur. Selektionskorrektur für Abiturnote berechnet mit geschätzter Gesamtvarianz; Selektionskorrektur für Auswahltest und Subtests gerechnet mit Varianz in ursprünglicher Gesamtpopulation (alle Testteilnehmer). * $p < .05$. ** $p < .01$. Fishers *z* wurde berechnet für die Unterschiede in den korrigierten Korrelationskoeffizienten zwischen Männern und Frauen, *df* (Fishers-*z*-Test) = 577.

Mithilfe von schrittweisen, vorwärts gerichteten multiplen Regressionsanalysen wurde überprüft, wie viel Varianz der Abschlussnote durch die Prädiktoren vorhergesagt wird. Bei einer ersten Analyse wurden die Abiturdurchschnittsnote, das Ergebnis im Auswahltest und das Geschlecht als Prädiktoren mit aufgenommen (s. Tabelle 4). Die Abiturnote klärt 15 % der Varianz auf. Durch die Aufnahme des Auswahltests in das Regressionsmodell werden zusätzliche 8 % Varianz aufgeklärt. Der Prädiktor Geschlecht wurde ausgeschlossen. Bei einer zweiten Regressionsanalyse wurden die Subtests in das Regressionsmodell aufgenommen. Von den Subtests tragen *Indizien, Fälle und Normen* sowie *Diagramme und Tabellen* mit einem signifikanten Beitrag zur Aufklärung der Varianz der Studienleistung bei (s. Tabelle 5). Das Geschlecht sowie die Subtests *Schlussfolgerungen* und *Oberbegriffe* wurden ausgeschlossen; sie tragen nicht signifikant zur Vorhersage der Abschlussnote bei.

Tab. 4: Schrittweise, vorwärts gerichtete multiple Regression mit Abiturdurchschnittsnote und Auswahltest.

Modell	Prädiktoren	R	R^2	R^2_{corr}	ΔR^2_{corr}	SE
1	Abi	.39	.16	.15		1.37
2	Abi Auswahltest	.49	.24	.23	.08	1.30

Anmerkungen: n = 579; *R* = multipler Korrelationskoeffizient; R^2 = Bestimmtheitsmaß; R^2_{corr} = korrigiertes Bestimmtheitsmaß; ΔR^2 = Zugewinn an aufgeklärter Varianz; *SE* = Standardfehler.

Tab. 5: Schrittweise, vorwärts gerichtete multiple Regression mit Abiturdurchschnittsnote und den Subtests des Auswahltests.

Modell	Prädiktoren	R	R^2	R^2_{corr}	ΔR^2_{corr}	SE
1	Abi	.39	.16	.15		1.37
2	Abi Indizien	.45	.21	.20	.05	1.33
3	Abi Indizien FuN	.48	.23	.23	.03	1.30
4	Abi Indizien FuN DuT	.49	.24	.24	.01	1.30

Anmerkungen: n = 579; FuN = *Fälle und Normen*; DuT = *Diagramme und Tabellen*; *R* = multipler Korrelationskoeffizient; R^2 = Bestimmtheitsmaß; R^2_{corr} = korrigiertes Bestimmtheitsmaß; ΔR^2 = Zugewinn an aufgeklärter Varianz; *SE* = Standardfehler.

3.3 Fairness

In Tabelle 6 werden die Ergebnisse der Vergleiche der Regressionsgeraden (mittels t-Tests für die Steigungsparameter, mittels F-Tests für die Standardschätzfehler) von Männern und Frauen für die verschiedenen Prädiktoren und das Kriterium aufgeführt. Die Steigungsparameter sowie die Standardschätzfehler der geschlechtsspezifischen Regressionsgeraden unterscheiden sich bei keinem der Prädiktoren signifikant voneinander. Für den Prädiktor Auswahltest, das schriftliche Auswahlverfahren und die Subtests *Diagramme und Tabellen* sowie *Schlussfolgerungen* ergeben sich auch keine signifikanten Unterschiede bezüglich der Ordinatenabschnitte der Regressionsgeraden, das heißt, sie sind nach Cleary (1968) als fair zu bewerten. Die Regressionsgeraden der Männer haben bei den Subtests *Indizien* und *Fälle und Normen* einen etwas höheren Ordinatenabschnitt als die der Frauen (d = 0.16 bis d = 0.26), bei dem Subtest *Oberbegriffe* ist es umgekehrt (d = - 0.37). Folglich werden die Leistungen der Frauen durch die Subtests *Indizien* und *Fälle und Normen* leicht überschätzt, während sie durch den Subtest *Oberbegriffe* unterschätzt werden. Diese Unterschiede scheinen sich auszugleichen, da sich für den Auswahltest insgesamt keine signifikanten Geschlechtsunterschiede ergeben.

Die Ergebnisse zur Überprüfung der Fairness nach Lawshe (1983) sind in Tabelle 7 aufgeführt. Die Differenzen zwischen den Prädiktoren und dem Kriterium LL.B.-Note unterscheiden sich für den Auswahltest, das schriftliche Auswahlverfahren sowie die Subtests *Indizien*, *Fälle und Normen* und *Oberbegriffe* nicht voneinander und sind somit als fair zu beurteilen. Die Abiturnote überschätzt die Studienleistung der Frauen leicht (d = 0.22), die Subtests *Schlussfolgerungen* und *Diagramme und Tabellen* unterschätzen die Studienleistung der Frauen (d = - 0.53 und d = - 0.47).

Tab. 6: Vergleich der Regressionsgeraden (Fairnessmodell von CLEARY, 1968).

	Kriterium: LL.B.						
Prädiktor	Parameter	Männer	Frauen	t / *F*	p	d	Bewertung
Abi	b	-1.56	-1.55	-0.03	.97	0.00	
	a	11.96	11.78	1.44	.15	0.13	Fair
	SE	1.36	1.38	*1.03*	.41	0.09	
Auswahltest	b	0.06	0.06	-0.26	.79	-0.02	
	a	5.95	5.90	0.44	.66	0.04	Fair
	SE	1.41	1.40	*1.01*	.47	0.09	
Abi und Auswahltest (1:2-Gewichtung)	b	0.28	0.30	-0.55	.96	-0.05	
	a	9.59	9.50	0.79	.43	0.07	Fair
	SE	1.34	1.32	*1.02*	.44	0.09	
Schlussflg	b	0.08	0.07	0.11	.92	0.01	
	a	8.57	8.75	-1.29	.20	-0.11	Fair
	SE	1.48	1.47	*1.01*	.47	0.09	
DuT	b	0.11	0.11	-0.15	.89	-0.01	
	a	8.36	8.44	-0.59	.56	-0.05	Fair
	SE	1.46	1.46	*1.01*	.47	0.09	
Indizien	b	0.12	0.13	-0.34	.74	-0.03	Überschätzung der Leistung der Frauen
	a	8.26	8.03	1.79	**.04**	**0.16**	
	SE	1.43	1.41	*1.03*	.41	0.09	
FuN	b	0.14	0.17	-0.63	.53	-0.05	Überschätzung der Leistung der Frauen
	a	8.20	7.83	2.98	**< .01**	**0.26**	
	SE	1.46	1.44	*1.02*	.44	0.09	
Oberbegr	b	0.08	0.05	0.62	.54	0.05	Unterschätzung der Leistung der Frauen
	a	8.20	8.75	-4.26	**< .01**	**-0.37**	
	SE	1.48	1.48	*1.00*	.51	0.09	

Anmerkungen: n = 579; $n_{Männer}$ = 378; n_{Frauen} = 201; LL.B. = Bachelor of Laws; Schlussflg = *Schlussfolgerungen*; DuT = *Diagramme und Tabellen*; FuN = *Fälle und Normen*; Oberbegr = *Oberbegriffe*; b = Steigung der Geraden; a = Konstante der Regressionsgleichung (Ordinatenabschnitt der Geraden); SE = Standardschätzfehler; *df* (*t*-Wert) = 575; *df* (*F*-Wert) = 376, 199. Positive Effektstärken bedeuten eine Überschätzung der Leistung der Frauen.

Tab. 7: Vergleich der Differenzen zwischen Prädiktoren und Kriterium von Männern und Frauen (Prüfung der Fairness nach LAWSHE, 1983).

	Männer		Frauen					
	M	SD	M	SD	t	p	d	Bewertung
Abi – LL.B.	-0.08	1.10	0.16	1.08	-2.55	**.01**	**0.22**	Überschätzung der Leistung der Frauen
Test – LL.B.	-0.01	1.16	0.01	1.15	-0.17	.87	0.02	Fair
Abi und Auswahltest (1:2) – LL.B.	-0.03	0.97	0.06	0.93	-1.11	.27	0.10	Fair
Schlussflg – LL.B.	0.24	1.27	-0.45	1.31	6.12	< **.01**	**-0.53**	Unterschätzung der Leistung der Frauen
DuT – LL.B.	0.20	1.26	-0.38	1.22	5.32	< **.01**	**-0.47**	Unterschätzung der Leistung der Frauen
Indizien – LL.B.	-0.05	1.21	0.10	1.17	-1.51	.13	0.13	Fair
FuN – LL.B.	0.00	1.28	0.00	1.17	.04	.97	0.00	Fair
Oberbegr – LL.B.	0.04	1.32	-0.08	1.35	.99	.32	-0.09	Fair

Anmerkungen: n = 579; $n_{Männer}$ = 378; n_{Frauen} = 201; LL.B. = Bachelor of Laws; Schlussflg = *Schlussfolgerungen*; DuT = *Diagramme und Tabellen*; FuN = *Fälle und Normen*; Oberbegr = *Oberbegriffe*; df (t-Wert) = 575. Positive Effektstärken bedeuten eine Überschätzung der Leistung der Frauen.

4 Diskussion

Der Auswahltest der BLS sagt die Leistung im Studium gut vorher und ist für beide Geschlechter prognostisch valide. Der Test ist zudem aufgrund der dargestellten Ergebnisse nach zwei Fairnessmodellen als fair gegenüber Männern und Frauen zu bewerten. Das heißt, dass die Studienleistung von Männern und Frauen durch das Ergebnis des Auswahltests weder über- noch unterschätzt wird. Der Auswahltest leistet neben der Abiturnote einen substantiellen Beitrag zur Prognose des Studienerfolgs. Auch das gesamte schriftliche Auswahlverfahren der BLS, bestehend aus den Kriterien Abiturdurchschnittsnote und Auswahltest mit der Gewichtung 1:2, sagt die Leistungen im Studium gut vorher und ist geschlechterfair.

Bezüglich der Abiturnote unterscheiden sich die Ergebnisse der beiden Berechnungsmethoden: Nach der Methode von Cleary (1968) ist die Abiturnote geschlechterfair, nach der Methode von Lawshe (1982) wird die Studienleistung von Männern unterschätzt, es handelt sich jedoch nur um einen kleinen Effekt. Die Abiturnote sagt den Studienerfolg bei Männern und Frauen gut vorher.

Insgesamt lässt sich feststellen, dass die Hinzunahme des Auswahltests das schriftliche Auswahlverfahren hinsichtlich Prognosekraft und Geschlechterfairness verbessert. Auf der Ebene der Aufgabengruppen des Auswahltests zeigt sich ein differenziertes Bild: Einige Aufgabengruppen scheinen Frauen leicht zu bevorzugen, andere Männer, während wiederum andere fair sind. Beide Methoden zur Fairness-Berechnung kommen deskriptiv zu ähnlichen Ergebnissen, weichen bei den Signifikanzaussagen aber zum Teil voneinander ab. So sind *Schlussfolgerungen* und *Diagramme und Tabellen* nach beiden Modellen deskriptiv unfair, aber nur nach Lawshe (1982) wird der Unterschied signifikant. Mit beiden Methoden ist es vermutlich möglich, größere Einschränkungen der Geschlechterfairness zu entdecken.

Die Ergebnisse dieser Studie stützen die Vermutung, dass eine Verwendung von komplexen, fachspezifischen Aufgabentypen zur Erfassung kognitiver Fähigkeiten die prognostische Validität und die Geschlechterfairness eines Auswahlverfahrens verbessern kann. Mit den Aufgabengruppen *Oberbegriffe*, *Schlussfolgerungen* und *Diagramme und Tabellen* wird die Leistung der Frauen tendenziell unterschätzt.

Zudem haben diese Aufgabengruppen nur eine mittlere Validität und tragen im Gesamtmodell nur geringfügig zur Prognose bei. Bei den Aufgabengruppen *Indizien* sowie *Fälle und Normen* wird die Leistung von Frauen tendenziell überschätzt. Der Fachbezug dieser Aufgabengruppen ist hoch und es handelt sich um Miniatursimulationen von Situationen im Studium. Im Anschluss an die durchgeführte Studie wurde der Auswahltest modifiziert: Die Aufgabengruppen *Oberbegriffe* sowie *Schlussfolgerungen* wurden entfernt, stattdessen wurde als neue Aufgabengruppe *Sprachgefühl* aufgenommen.

Zukünftige Studien sollten die Eigenschaften verschiedener Aufgabentypen weiter untersuchen und prüfen, ob sich Fachbezug und Komplexität positiv auf Validität und Geschlechterfairness auswirken.

Die vorliegende Studie erweitert das Verständnis von fachspezifischen SFTs und deren Geschlechterfairness. Sie untersucht mit einer großen Stichprobe einen Auswahltest für Rechtswissenschaften und dessen Prognosekraft für Abschlussnoten. Sie verwendet zwei Methoden zur Fairnessberechnung und ermittelt die Fairness für das gesamte schriftliche Auswahlverfahren sowie für dessen Bestandteile. Sie zeigt außerdem auf, wie ein Auswahlverfahren evaluiert und im Anschluss optimiert werden kann.

Die Einschränkung der Studie besteht darin, dass ein Auswahltest für Rechtswissenschaften mit den Daten einer Hochschule untersucht wurde und die Ergebnisse nicht ohne weiteres auf andere Hochschulen und Studienfächer übertragen werden können. Zudem wäre es wünschenswert gewesen, auch die Ergebnisse des mündlichen Auswahlverfahrens zu untersuchen, um eine Aussage über das gesamte Auswahlverfahren treffen zu können. Weiterhin wurden nur Abschlussnoten als Erfolgskriterium untersucht; es gab keine Analyse der Studiendauer oder Studienabbrüche, unter anderem, da die Abbruchquote an der BLS sehr niedrig ist.

Besonders wichtig ist es nach Auffassung der Autoren, zukünftig nicht nur zu untersuchen, ob ein Test fair ist, sondern genauer zu ergründen, warum ein (Sub-)Test fair oder unfair ist, und daraus abzuleiten, wie Tests zukünftig gestaltet werden müssen, damit sie die bestmögliche Validität und Fairness haben.

5 Literaturverzeichnis

Alexander, R. A., Alliger, G. M. & Hanges, P. J. (1984). Correcting for range restriction when the population variance is unknown. *Applied Psychological Measurement, 8*, 431-437.

Bucerius Law School (2018). Broschüre zur Vorbereitung auf das schriftliche Auswahlverfahren. https://www.law-school.de/jurastudium/bewerbung/auswahlverfahren/schriftliches-auswahlverfahren/

Cleary, T. A. (1968). Test bias: prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, *5*, 115-124.

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

Deidesheimer Kreis (Amelang, M., Bartussek, D., Brackmann, H.-J., Egli, H., Haase, K., Hinrichsen, K., Klauer, K.J., Kornadt, H.-J., Michel, L. & Trost, G.) (1997). *Hochschulzulassung und Studieneignungstests. Studienfeldbezogene Verfahren zur Feststellung der Eignung für Numerus-clausus- und andere Studiengänge.* Göttingen, Zürich: Vandenhoeck & Ruprecht.

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

Fischer, F., Schult, J. & Hell, B. (2015). Unterschätzung der Studienleistungen von Frauen durch Studierfähigkeitstests: Erklärbar durch Persönlichkeitseigenschaften? *Diagnostica, 61,* 34-46. https://doi.org/10.1026/0012-1924/a000120

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

Hancock, T. (1999). The gender difference: validity of standardized admission tests in predicting MBA performance. *Journal of Education for Business, 75,* 91-94.

Hänsgen, K.-D. & Spicher, B. (2001). *EMS – Eignungstest für das Medizinstudium in der Schweiz. Vorhersage des Prüfungserfolges*. Bericht 7. Fribourg: ZTD.

Hell, B. & Schuler, H. (2005). Verfahren der Studierendenauswahl aus Sicht der Bewerber. *Empirische Pädagogik, 19,* 361-376.

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

Herde, C. N., Stegt, S. J. & Preckel, F. (2016). Verfahren der Studierendenauswahl für Masterstudiengänge aus Sicht der Bewerber. *Zeitschrift für Arbeits- und Organisationspsychologie, 60*(1), 145-161.

Lawshe, C. H. (1983). A simplified approach to the evaluation of fairness in employee selection procedures. *Personnel Psychology*, *36*, 601-608.

Lienert, G. A. (1967). *Testaufbau und Testanalyse* (2. Aufl.). Weinheim, Berlin: Verlag Julius Beltz.

Liepmann, D., Beauducel, A., Brocke, B. & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000 R* (IST 2000 R). Manual (2., erweiterte und überarbeitete Aufl.). Göttingen: Hogrefe.

Nauels, H.-U. & Meyer, M. (1997). Untersuchungen zur Vorhersagekraft des TMS: differentielle Aspekte der Studienerfolgsprognose und Testfairneß. In G. Trost (Hrsg.), *Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 21. Arbeitsbericht* (S. 76-108). Bonn: Institut für Test- und Begabungsforschung.

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.

Schult, J., Hell, B., Päßler, K. & Schuler, H. (2013). Sex-specific differential prediction of academic achievement by german ability tests. *International Journal of Selection and Assessment, 21,* 130-134. https://doi.org/10.1111/ijsa.12023

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. https://doi.org/10.1037/0022-0663.85.4.710

Young, J. W. & Kobrin, J. L. (2001). *Differential validity, differential prediction, and college admission testing: a comprehensive review and analysis*. (College Board No. 2001-6). New York, NY: College Entrance Examination Board.

Autoren

Luisa BERGHOLZ || ITB Consulting GmbH ||
Koblenzer Straße 77, D-53177 Bonn

www.itb-consulting.de

luisa.bergholz@itb-consulting.de

Dr. Stephan Josef STEGT || ITB Consulting GmbH ||
Koblenzer Straße 77, D-53177 Bonn

www.itb-consulting.de

stephan.stegt@itb-consulting.de