C.3 Ökonomie und Fairness von Constructed-Response-Items in E-Assessments

Norbert Pengel, Patrick Hawlitschek, Marios Karapanos
Universität Leipzig

1 Problemstellung

Das Testen kognitiver Fähigkeiten ist ein Standardproblem in der Leistungsdiagnostik. Typische Anwendungsfelder sind Lernstandsüberprüfungen, Zulassungs- und Auswahlverfahren an Schulen und Hochschulen, aber auch Eignungs- und Einstellungsverfahren im betrieblichen Kontext. Sollen nicht nur einige wenige Personen, sondern größere Kohorten getestet werden, kommen aus testökonomischen Gründen vermehrt computerbasierte Tests (E-Assessments) und Aufgaben mit geschlossenem Antwortformat (Selected-Response, SR) zum Einsatz. Auf diese Weise lassen sich Tests automatisch auswerten, was gegenüber papierbasierten Tests den Testaufwand erheblich reduziert und eine zeitnahe Rückmeldung der Ergebnisse an die getesteten Personen erlaubt (Michel, Goertz, Radomski, Fritsch, & Baschour, 2015). Neben der guten Testökonomie besitzen SR-Tests auch aus psychometrischer Sicht Vorteile. So lassen sie sich nicht nur objektiver auswerten, sondern auch zeitsparender beantworten, wodurch eine größere Zahl an Aufgaben bei gleicher Testdauer gestellt werden kann Lindner, Strobel, & Köller, 2015). Dennoch werden SR-Tests insbesondere an Hochschulen häufig als besonders rigide Prüfungsform wahrgenommen (Kubinger, 2014). Eine wiederkehrend zu beobachtende Strategie scheint deshalb – wohl auch um die Akzeptanz dieser Prüfungsform zu erhöhen – die Ergänzung von E-Assessments um Freitextaufgaben (Constructed-Response, CR) zu sein. Hochschulprüfungen entscheiden über den Zugang zu erstrebenswerten Gütern einer Gesellschaft (Huinink & Schröder, 2014) und ziehen berufliche Auswahlentscheidungen nach sich (*Rekrutierungsfunktion*; Tsarouha, 2019). Vor dem Hintergrund der grundgesetzlich geregelten Berufswahlfreiheit (Artikel 12, GG) ergibt sich die berechtigte Forderung nach einer hohen diagnostischen Güte der eingesetzten Tests. Gleichzeitig erfordern die institutionellen Rahmenbedingungen an Hochschulen ökonomische Testmethoden. Aktuell fehlt es an Arbeiten, die den Verlust an Testökonomie durch Hinzunahme von CR-Items quantifizieren und den möglichen Gewinn an diagnostischer Güte zueinander ins Verhältnis setzen. Zudem weisen Schulleistungsstudien (Lafontaine & Monseur, 2009; Lissitz, Hou, & Slater, 2012; Reardon, Kalogrides, Fahle, Podolsky, & Zárate, 2018) und Untersuchungen aus dem Hochschulkontext (Arthur & Everaert, 2012) auf geschlechterdifferenzielle Effekte verschiedener Itemformate hin, die zu einem Problem für die Testfairness werden können. Beide Forschungsfragen adressiert der vorliegende Beitrag anhand einer empirischen Analyse von Daten einer E-Klausur, die die Abschlussprüfung eines erziehungswissenschaftlichen Moduls im universitären Lehramtsstudium bildet.

2 Testökonomie und Testfairness

Ökonomie und Fairness gelten als messtheoretische Nebengütekriterien eines Tests (Moosbrugger & Kelava, 2012; Schmidt-Atzert & Amelang, 2012). Als solche werden sie in der Literatur zumeist nachgeordnet behandelt, obwohl sie in der angewandten Diagnostik, etwa im Kontext von Hochschulprüfungen oder in der Personalauswahl, von hoher praktischer Bedeutung sind. Ein Test gilt dann als ökonomisch, „wenn er, gemessen am diagnostischen Erkenntnisgewinn, relativ wenig finanzielle und zeitliche Ressourcen beansprucht" (Moosbrugger & Kelava, 2012, S. 21). E-Assessments tragen insgesamt zur Verbesserung der Testökonomie bei. Sie erlauben nicht nur die automatische Auswertung von SR-Items. Auch CR-Items lassen sich durch das einheitliche Schriftbild oft effizienter auswerten (Stieler, 2011). Die Bewertung muss allerdings noch weitestgehend manuell erfolgen. Algorithmische Verfahren zur automatisierten Bewertung von Freitexten (*Automated Essay Scoring*) befinden sich seit vielen Jahren in der Entwicklung und erzielen in ausgewählten Anwendungen bereits ein erstaunliches Maß an Übereinstimmung mit menschlichen Ratern ($r > .90$ in Alikaniotis, Yannakoudakis, & Rei, 2016; $r > .70$ in Rupp et al. 2019). Sie erfordern allerdings einen großen Umfang an Trainingsdaten und sind noch nicht ausreichend robust für den unüberwachten Feldeinsatz. Eine Nutzung an deutschen Hochschulen erscheint auf absehbare Zeit ausgeschlossen.

Da bei Verwendung von SR-Items Testpersonen die Antwort nicht selbst verschriftlichen, sondern nur aus vorgegebenen Antworten auswählen, erfassen SR-Items weniger konstruktfremde Varianz und liefern aufgrund der eingesparten Schreibzeit mehr diagnostische Information bei gleicher Testdauer (Lindner et al., 2015; Wan & Henley, 2012, Lukhele, Thissen, & Wainer, 1994). In kombinierten Tests zeigen umfangreiche CR-Items (Essays) eine geringere prädiktive Validität (Breland, Kubota, & Bonner, 1999; Bridgeman, 1991; Norris, Oppler, Kuang, Day, & Adams, 2006). SR-Tests gelten deshalb gegenüber CR-Tests als das ökonomischere und validere Verfahren. Wie hoch der Gewinn an Testökonomie im Rahmen von Hochschulprüfungen ausfällt, muss noch anhand von Praxisdaten ermittelt werden (Lindner et al., 2015).

Fairness ist bei einem Test dann gegeben, „wenn die resultierenden Testwerte zu keiner systematischen Benachteiligung bestimmter Personen aufgrund ihrer Zugehörigkeit zu ethnischen, soziokulturellen oder geschlechtsspezifischen Gruppen führen" (Moosbrugger & Kelava, 2012, S. 24). Gruppenunterschiede allein sind jedoch noch kein Beleg für einen unfairen Test (Zieky, 2016). Eine Reihe von Arbeiten hat sich in der Vergangenheit mit der Frage befasst, ob bestimmte Itemformate für einzelne Gruppen leichter oder schwerer zu lösen sind. In der Zusammenschau zeigt sich, dass weibliche Testpersonen typischerweise bei CR-Items im Vorteil sind und männliche

Testpersonen besser bei SR-Items abschneiden (Arthur & Everaert, 2012; Gamer & Engelhard Jr., 1999; Reardon et al., 2018). Letztgenannter Effekt tritt allerdings weniger stabil auf (Lissitz et al., 2012; Liu & Wilson, 2009). Je nach Domäne und Zusammenstellung des Tests kann sich dadurch der Leistungsunterschied zwischen männlichen und weiblichen Testpersonen verringern oder vergrößern.

Als Erklärung für die besseren Leistungen von weiblichen Testpersonen bei CR-Items kommen vor allem Unterschiede in verbalen Fähigkeiten in Frage. Stärker als SR-Items erfassen CR-Items auch verbale Fähigkeiten der Testperson. Diese sind in der Regel nicht Gegenstand der eigentlichen Messung, lassen sich aber in der praktischen Bewertung oft nur schwer vom tatsächlich zu messenden Konstrukt trennen (Lindner et al., 2015). Grundlegend wird angenommen, dass sich Männer und Frauen in den meisten kognitiven Fähigkeiten nicht (nennenswert) unterscheiden (*Gender Similarities Hypothesis*; Hyde, 2005; auch Hedges & Novell, 1995). Verbale Fähigkeiten stellen aber eine wiederkehrend diskutierte Ausnahme dar. Nach einer Meta-Analyse von Hyde und Linn (1988) bestehen die größten Unterschiede im sprachlichen Ausdruck ($d = 0.33$), beim Lösen von Anagrammen ($d = 0.22$) und in der verbalen Grundfähigkeit (*general verbal ability*, $d = 0.20$). Neuere Untersuchungen weisen aber auch insbesondere auf Unterschiede in der Lese- und Schreibfähigkeit hin ($d = 0.23$ bzw. $d = 0.46$ in Reynolds, Scheiber, Hajovsky, Schwartz, & Kaufman, 2015; Olson et al., 2013; Scheiber, Reynolds, Hajovsky, & Kaufman, 2015; Camarata & Woodcock, 2006; Reilly, Neumann, & Andrews, 2018). Die Effekte treten stärker bei Tests unter Zeitvorgabe hervor (Camarata & Woodcock, 2006), ein für Prüfungssituationen typisches Merkmal. Auch in Schulleistungsuntersuchungen erzielen Mädchen auffallend stabil höhere Leistungen im Lesen und Schreiben als Jungen (Stoet & Geary, 2013; Naumann, Artelt, Schneider & Stanat, 2010; Fischer, Schult & Hell, 2013).

Testfairness bezieht sich immer auf eine konkrete Testsituation und ist deshalb kein Testmerkmal im engeren Sinne. Da die Aufnahme eines Hochschulstudiums mit einer hohen Eingangsselektion verbunden ist, bleibt zu klären, ob mögliche Geschlechtereffekte – insbesondere in zulassungsbeschränkten Studiengängen – bereits auf diese Weise nivelliert werden oder ob sie in relevantem Ausmaß bestehen bleiben.

3 Methode

Der vorliegende Beitrag untersucht die Auswirkungen von CR-Items auf Ökonomie und Fairness von E-Assessments. Dazu werden Daten einer computergestützten Klausur (E-Klausur) analysiert. Sie bildet die Modulprüfung eines bildungswissenschaftlichen Moduls der Lehramtsstudiengänge an der Universität Leipzig. Die Klausur enthält 29

SR-Items und zwei CR-Items, von denen die Studierenden eines zur Beantwortung auswählen. Die Bewertung der CR-Items erfolgt durch zwei Prüfende. Das arithmetische Mittel beider Bewertungen ergibt den Punktwert für das Item. Zur Schätzung des Bewertungsaufwands werden Logdaten des Testsystems analysiert, das die Bewertung der CR-Items protokolliert. Für jede abgeschlossene Bewertung wird im System ein Speicherzeitpunkt angelegt. Die Bewertungsdauer wird aus der Differenz zwischen den Speicherzeitpunkten von zwei aufeinander folgenden Bewertungen desselben Prüfenden bestimmt. Durch dieses Vorgehen kann zwar nicht die Bewertungsdauer der jeweils ersten Bewertung einer Serie ermittelt werden, da der Anfangszeitpunkt fehlt. Weil davon auszugehen ist, dass die benötigte Zeit für eine Bewertung unabhängig von der Position einer Aufgabe innerhalb einer Serie ist, ist die Genauigkeit der Schätzung durch die fehlenden Daten aber kaum beeinträchtigt. Um die Schätzung auf mögliche Arbeitsunterbrechungen zu korrigieren, wird eine 5%-Trimmung der Daten am oberen Ende der Verteilung vorgenommen. Die Klausurdatensätze enthalten keine Angaben zum Geschlecht der geprüften Personen, weshalb eine algorithmische Klassifikation auf Basis der Vornamen vorgenommen wird (Wais, 2016). Personen mit Vornamen, die keine eindeutige Klassifikation als männlich oder weiblich zulassen, werden aus der Analyse entfernt. Die Auswertung beschränkt sich auf die Daten der vier großen Lehramtsstudiengänge (Grundschule, Gymnasium, Oberschule, Sonderpädagogik). Für diese liegen ausreichend Fallzahlen für eine robuste Analyse vor. In Summe verbleiben so aus 17 untersuchten Parallelprüfungen 757 getestete Personen (527 weiblich).

Die Fairness eines Tests lässt sich prinzipiell auf zwei Arten bestimmen. Interne Verfahren prüfen, ob psychometrische Eigenschaften eines Tests zwischen demographischen Gruppen variieren (Meade & Tonidandel, 2010). Mögliche Ansätze sind dabei die differentielle Bestimmung von Itemfunktionen auf der Grundlage der Item-Response-Theorie oder die Prüfung auf Invarianz des Messmodells anhand konfirmatorischer Faktorenanalysen (Meade & Tonidandel, 2010). Interne Verfahren gelten als Mittel der Wahl, lassen sich aber nicht uneingeschränkt auf Hochschulprüfungen anwenden. Zum einen messen Hochschulprüfungen in der Regel kein eindimensionales Fähigkeitskonstrukt, sondern oft sehr heterogene Kompetenzen und Wissensbestände (Marcus, 2015). Zum anderen erfordern diese Verfahren sehr große Samples je Item, die selbst in Massenstudiengängen oft kaum zu erzielen sind. Die Alternative stellen externe Verfahren dar. Sie überprüfen die Fairness anhand statistischer Zusammenhänge mit (externen) Kriterien (Meade & Tonidandel, 2010) wie bspw. Studienerfolg oder Jobperformance. Wäre ein Test unfair, so würde er das externe Kriterium für die durch den Test benachteiligte Gruppe systematisch unterschätzen. Ein klassischer und häufig genutzter Ansatz stammt von Cleary (1968), der mittels hierarchischer Regressionsanalysen die Angemessenheit einer gemeinsamen

Regressionsgerade für alle in Frage stehenden Gruppen untersucht (Schmidt-Atzert & Amelang, 2012). Der Test ist fair, wenn es keine signifikanten Unterschiede zwischen den Konstanten und Anstiegen der Regressionsgeraden aller Gruppen gibt. Als Kriterium wird im vorliegenden Fall der Punktwert aus dem geschlossenen Teil der Prüfung (29 SR-Items) genutzt. Zwar stellt dieser kein echtes externes Kriterium dar. Es kann aber angenommen werden, dass er wegen der hohen Validität von SR-Tests einen zuverlässigen Schätzer für das wahre Fähigkeitsniveau der Testperson darstellt.

4 Ergebnisse

4.1 Ökonomie

Durch das oben beschriebene Verfahren zur Bestimmung der Bewertungszeiten verbleiben 573 gültige Fälle in der Analyse. Im Mittel dauert die Bewertung einer Freitextlösung 4.36 Minuten (SD = 3.93 Minuten). Die Verteilung ist unimodal und rechtsschief (*Skewness* = 2.12). Die durchschnittliche Länge eines Lösungstexts im Datensatz beträgt 2106 Zeichen mit Leerzeichen (ZML; SD = 817.6). Die mittlere Lesegeschwindigkeit akademisch vorgebildeter Personen kann für mittelschwere Texte mit etwa 1500 ZML pro Minute angenommen werden (Musch & Röseler, 2011). Eine Bewertung dauert damit etwa dreimal so lang wie das reine Lesen der Freitextlösung. Müssen, wie im vorliegenden Fall, pro Semester ca. 400 bis 500 Klausuren bewertet werden, verursacht die Integration einer einzigen Freitextaufgabe einen Zusatzaufwand von ca. 58 bis 72 Personenstunden für Erst- und Zweitkorrektur. Nicht berücksichtigt sind dabei organisatorische Aufwände, z.B. für die Verteilung der Aufgaben zwischen verschiedenen Prüfenden und Abstimmungsbedarf zwischen Erst- und Zweitprüfenden. Ausgehend vom aktuellen DFG-Personalmittelsatz von 6000 Euro pro Monat (Postdoc oder vergleichbar; DFG, 2019), entstehen auf diese Weise Personalkosten zwischen 2500 und 3100 Euro (20 Arbeitstage im Monat abzüglich 2.5 Urlaubstage nach Tarifvertrag der Länder).

Über alle Aufgaben betrachtet, korrelieren die Punktwerte von SR- und CR-Items zu $r = .45$, $p < .001$, 95% CI [.40, .51]. Zum Vergleich: Die mittlere Trennschärfe (Median aller Trennschärfekoeffizienten) aller SR-Items liegt bei $r = .27$, 95% CI [.25, .29] (BCa Bootstrap, 5000 Samples). Die CR-Items liefern damit nicht ausschließlich redundante Informationen über die Testpersonen, sondern tragen eigenständig zum diagnostischen Erkenntnisgewinn bei.

4.2 Fairness

Im ersten Schritt werden männliche und weibliche Testpersonen in Test und Kriterium auf Mittelwertunterschiede untersucht (Meade & Tonidandel, 2010). Aus Seminarbeobachtungen ist bekannt, dass die verschiedenen Lehramtsstudiengänge Studierende mit unterschiedlichen Fähigkeits- und Persönlichkeitsprofilen rekrutieren.

Tabelle 1: Mittelwerte und Standardabweichungen für CR- und SR-Items

		CR		SR	
	N	***M***	***SD***	***M***	***SD***
Männer					
Grundschule	23	9.13	2.82	77.61	14.65
Gymnasium	130	9.64	3.43	80.53	12.38
Oberschule	51	7.37	3.49	72.63	11.95
Sonderpädagogik	26	8.92	2.45	83.46	11.15
Frauen					
Grundschule	171	10.16	3.38	83.82	11.82
Gymnasium	168	9.88	3.61	81.83	13.04
Oberschule	63	8.56	3.14	72.81	12.78
Sonderpädagogik	125	9.71	3.39	82.30	11.78

Da das Geschlechterverhältnis über die vier betrachteten Lehramtsstudiengänge variiert ($\chi^2 = 79.66$, $p < .001$), werden die Punktwerte der CR-Items und SR-Items mittels zweifaktorieller ANOVA (Geschlecht x Studiengang) auf Mittelwertunterschiede getestet. Varianzgleichheit kann in beiden Analysen angenommen werden (Levene-Test, $F = 1.203$, $p = .299$ bzw. $F = 0.815$, $p = .575$). Die Analyse zeigt signifikante Haupteffekte für Geschlecht ($F (1, 749) = 6.308$, $p = .012$, partielles $\eta^2 = 0.008$) und Studiengang ($F (3, 749) = 7.899$, $p < .001$, partielles $\eta^2 = 0.031$) auf die Punktwerte der CR-Items, aber keine Interaktion ($F < 1$).

Tabelle 2: Regression der SR-Bewertung auf CR-Bewertung und Studiengang

Model		*b*	*SE*	*β*	*t*	*p*
0	Konstante	60.248	1.426		42.254	< .001
	CR	1.555	0.120	0.419	12.925	< .001
	Gymnasium	5.846	1.252	0.223	4.670	< .001
	Sonderpädagogik	7.331	1.404	0.228	5.222	< .001
	Grundschule	7.229	1.344	0.247	5.380	< .001
1	Konstante	60.493	1.510		40.052	< .001
	CR	1.550	0.121	0.418	12.840	< .001
	Gymnasium	5.849	1.252	0.223	4.671	< .001
	Sonderpädagogik	7.211	1.425	0.225	5.059	< .001
	Grundschule	7.086	1.375	0.242	5.152	< .001
	Geschlecht (1 = männlich)	-0.464	0.940	-0.017	-0.494	0.621
2	Konstante	59.609	1.719		34.672	< .001
	CR	1.632	0.143	0.440	11.445	< .001
	Gymnasium	5.991	1.259	0.228	4.758	< .001
	Sonderpädagogik	7.302	1.428	0.227	5.114	< .001
	Grundschule	7.149	1.376	0.244	5.193	< .001
	Geschlecht (1 = männlich)	2.122	2.580	0.076	0.822	0.411
	Geschlecht x CR	-0.281	0.262	-0.100	-1.076	0.282

Anmerkungen. Referenzgruppe: Lehramt Oberschule

Während Lehramtsstudierende für Gymnasium, Grundschule und Sonderpädagogik vergleichbare Punktwerte erzielen (Tukey-Test, $p > .05$), schneiden Studierende für das Lehramt Oberschule gegenüber Gymnasium ($p < .001$, $d = 0.52$), Grundschule ($p = .004$, $d = 0.50$) und Sonderpädagogik ($p = .027$, $d = 0.41$) schlechter ab. Im geschlossenen Teil zeigt sich ebenfalls ein signifikanter Haupteffekt für Studiengang ($F (3, 749) = 15.502$, $p < .001$, partielles $\eta^2 = 0.058$), aber nicht für Geschlecht ($F (1, 749) = 1.929$, $p = .165$) und auch keine Interaktion ($F < 1$). Auch hier erzielen Studierende für das Lehramt Oberschule schlechtere Ergebnisse als Studierende in den Studiengängen Gymnasium ($p < .001$, $d = 0.67$), Grundschule ($p < .001$, $d = 0.65$) und Sonderpädagogik ($p < .001$, $d = 0.85$). Wegen der teilweise variierenden Leistung zwischen den einzelnen Lehramtsstudiengängen wird das Merkmal Studiengang in der regressionsanalytischen Überprüfung der Testfairness als zusätzlicher Prädiktor berücksichtigt.

Das Nullmodell zeigt einen statistisch signifikanten Zusammenhang zwischen Modellprädiktoren und Kriterium (F (4, 750) = 59.485, $p < .001$, adj. $R^2 = .24$). Die Hinzunahme der Variable Geschlecht (Model 1; F (1, 749) = 0.244, $p = .621$) und des Interaktionsterms Geschlecht x CR (Model 2; F (1, 748) = 1.159, $p = .282$) führen zu keiner signifikant besseren Modellanpassung (siehe Tabelle 2). Trotz der im Durchschnitt schlechteren Leistungen von männlichen Studierenden bei CR-Items ($d = .24$), liegen damit keine Hinweise auf einen Testbias vor.

5 Diskussion und Ausblick

Vor dem Hintergrund steigender Studierendenzahlen sollten vorhandene Ressourcen so eingesetzt werden, dass Hochschulprüfungen die eingangs beschriebenen Funktionen erfüllen und eine hohe diagnostische Güte aufweisen. CR-Items können in Verbindung mit SR-Items zu einem diagnostischen Mehrwert führen, sie erzeugen jedoch auch unter den Bedingungen von E-Assessments einen hohen Bewertungsaufwand. Gemessen an der Stärke des korrelativen Zusammenhangs zwischen SR- und CR-Testergebnissen erscheint dieser Mehraufwand für teilnehmerstarke Prüfungssituationen überdenkenswert. Statt beide Aufgabentypen in Prüfungen zu kombinieren, könnten die Personalressourcen, die sonst für die manuelle Korrektur der Freitextaufgaben eingesetzt werden müssen, möglicherweise sogar sinnvoller für die Erstellung und Qualitätssicherung geschlossener Aufgaben genutzt werden (Pengel, Thor, Seifert, & Wollersheim, 2017). SR-Tests stellen nur dann ein valides Prüfungsinstrument dar, wenn sie nach entsprechenden Gütekriterien konstruiert werden.

Kommen CR-Items in E-Assessments an Hochschulen zum Einsatz, so scheinen daraus keine negativen Folgen für die Testfairness zu resultieren. Für eine E-Klausur der universitären Lehrerbildung konnte dieser Beitrag zeigen, dass die genutzten CR-Items keine geschlechterdifferenziellen Effekte aufweisen. Ob es sich dabei um einen robusten und generalisierbaren Befund im Kontext von Hochschulprüfungen handelt, wird Gegenstand zukünftiger Analysen sein.

Literatur

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. arXiv:1606.04289 [cs].

Arthur, N., & Everaert, P. (2012). Gender and performance in accounting examinations: Exploring the impact of examination format. Accounting Education, 21(5), 471–487.

Breland, H. M., Kubota, M. Y., & Bonner, M. W. (1999). The performance assessment study in writing: Analysis of the SAT II: Writing Subject Test. New York: College Board Publications.

Bridgeman, B. (1991). Essays and multiple-choice tests as predictors of college freshman GPA. Research in Higher Education, 32(3), 319–332.
Camarata, S., & Woodcock, R. (2006). Sex differences in processing speed: Developmental effects in males and females. Intelligence, 34(3), 231–252.
Cleary, T. A. (1968). Test bias: Prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement, 5, 115–124.
DFG (2019). Personalmittelsätze der DFG für das Jahr 2019. DFG-Vordruck 60.12 – 01/19. Bonn: Deutsche Forschungsgemeinschaft.
Gamer, M., & Engelhard Jr., G. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. Applied Measurement in Education, 12(1), 29–51.
Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. Science, 269(5220), 41–45.
Huinink, J. & Schröder, T. (2014). Sozialstruktur Deutschlands. (2. Auflage). Konstanz, München: UVK.
Hyde, J., & Linn, M. (1988). Gender differences in verbal ability: A meta-analysis. Psychological Bulletin, 104(1), 53–69.
Hyde, J. S. (2005). The gender similarities hypothesis. American Psychologist, 60(6), 581–592
Kubinger, K. D. (2014). Gutachten zur Erstellung „gerichtsfester" Multiple-Choice-Prüfungsaufgaben. Psychologische Rundschau, 65(3), 169–178.
Lafontaine, D., & Monseur, C. (2009). Gender gap in comparative studies of reading comprehension: To what extent do the test characteristics make a difference? European Educational Research Journal, 8(1), 69–79.
Lindner, M. A., Strobel, B., & Köller, O. (2015). Multiple-Choice-Prüfungen an Hochschulen? Zeitschrift für Pädagogische Psychologie, 29(3–4), 133–149.
Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. Journal of Applied Testing Technology, 13(3).
Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. Applied Measurement in Education, 22(2), 164–184.
Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. Journal of Educational Measurement, 31(3), 234–250.
Marcus, B. (2015). Multiple-Choice-Prüfungsaufgaben in der Psychologie. Psychologische Rundschau, 66(3), 166–170.
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(2), 192–205.
Michel, L. P., Goertz, L., Radomski, S., Fritsch, T., & Baschour, L. (2015). Digitales Prüfen und Bewerten im Hochschulbereich. Berlin: Hochschulforum Digitalisierung.

Moosbrugger, H. & Kelava, A. (2012). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In Moosbrugger, H. & Kelava, A. (Hrsg.). Testtheorie und Fragebogenkonstruktion (S. 7–26). Berlin, Heidelberg: Springer.

Musch, J., & Rösler, P. (2011). Schnell-Lesen: Was ist die Grenze der menschlichen Lesegeschwindigkeit? In M. Dresler (Hrsg.), Kognitive Leistungen: Intelligenz und mentale Fähigkeiten im Spiegel der Neurowissenschaften (S. 89–106). Heidelberg: Spektrum.

Naumann, J., Artelt, C., Schneider, W. & Stanat, P. (2010). Lesekompetenz von PISA 2000 bis PISA 2009. In E. Klieme, C. Artelt, J. Hartig, N. Jude, O. Köller, M. Prenzel, W. Schneider, & P. Stanat (Hrsg.). PISA 2009. Bilanz nach einem Jahrzehnt (S. 23–72). Münster: Waxmann.

Norris, D., Oppler, S., Kuang, D., Day, R., & Adams, K. (2006). The College Board SAT® Writing Validation Study: An assessment of predictive and incremental validity. College Board Publications: New York.

Olson, R. K., Hulslander, J., Christopher, M., Keenan, J. M., Wadsworth, S. J., Willcutt, E. G., … DeFries, J. C. (2013). Genetic and environmental influences on writing and their relations to language and reading. Annals of Dyslexia, 63(1), 25–43.

Pengel, N., Thor, A., Seifert, P., & Wollersheim, H. W. (2017). Digitalisierte Hochschuldidaktik: Technologische Infrastrukturen für kompetenzorientierte E-Assessments – In: C. Igel (Hrsg.): Bildungsräume. Proceedings der 25. Jahrestagung der GMW (S. 232–238). Münster, New York: Waxmann.

Reardon, S. F., Kalogrides, D., Fahle, E. M., Podolsky, A., & Zárate, R. C. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eighth Grades. Educational Researcher, 47(5), 284–294.

Reilly, D., Neumann, D. L., & Andrews, G. (2018). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). American Psychologist, 74(4), 445–458.

Reynolds, M. R., Scheiber, C., Hajovsky, D. B., Schwartz, B., & Kaufman, A. S. (2015). Gender differences in academic achievement: Is writing an exception to the gender similarities hypothesis? The Journal of Genetic Psychology, 176(4), 211–234.

Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S., & Köller, O. (2019). Automated essay scoring at scale: A case study in Switzerland and Germany (Nr. RR–19–12). Educational Testing Service.

Schmidt-Atzert, L. & Amelang, M. (2012). Psychologische Diagnostik. (5. Auflage). Berlin, Heidelberg: Springer.

Scheiber, C., Reynolds, M. R., Hajovsky, D. B., & Kaufman, A. S. (2015). Gender differences in achievement in a large, nationally representative sample of children and adolescents. Psychology in the Schools, 52(4), 335–348.

Stieler, J. F. (2011). Validität summativer Prüfungen. Überlegungen zur Gestaltung von Klausuren. Bielefeld: Janus Presse.

Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within- and across-nation assessment of 10 Years of PISA data. PLoS ONE, 8(3).

Tsarouha, E. (2019). Prüfungspraktiken an deutschen Hochschulen. Eine empirische Studie zu systematischen Einflussgrößen auf die Notengebung in Abschlussprüfung. Wiesbaden: Springer VS.

Wais, K. (2016). Gender prediction methods based on first names with genderizeR. The R Journal, 8(1), 17–37.

Zieky, M. J. (2016). Developing fair tests. In S. Lane, M. R. Raymond, & T. M. Haladyna (Hrsg.), Handbook of Test Development (Second Edition, S. 81–99). New York: Routledge.