dgp
Deutsche Gesellschaft
für Personalwesen e.V.

Welchen Nutzen bringt Proctoring von Online-Tests?

■ DR. ANNA-LENA JOBMANN & AMELIE KLEINMANNS

*Immer mehr Online-Tests werden digital mittels Proctoring beaufsichtigt. Doch welchen konkreten Nutzen hat Online-Proctoring, und wiegt dieser Nutzen die mit Proctoring verbundenen Kosten auf Seiten von Bewerber*innen und Organisationen auf? Kann der Einsatz von Proctoring bei Online-Tests die Betrugsrate signifikant verringern? Unsere Recherchen und Analysen zeigen: Ja!*

ISSN 2942-1403

Die digitale Beaufsichtigung von Online-Tests (Proctoring) ist seit Beginn der Pandemie zunehmend verbreitet. Im Beitrag von Jobmann & Kleinmanns (2023) zum Thema „Proctoring – Ja oder nein?“ werden die grundsätzlichen Vor- und Nachteile der digitalen Beaufsichtigung von Online-Tests präsentiert und Empfehlungen zur Spezifikation gegeben. Doch welchen konkreten Nutzen hat Online-Proctoring, und wiegt dieser Nutzen die mit Proctoring verbundenen Kosten auf Seiten von Bewerber*innen und Organisationen auf? Kann der Einsatz von Proctoring bei Online-Tests die Betrugsrate signifikant verringern? Zunächst präsentieren wir in diesem Beitrag den Stand der Forschung zur Wirksamkeit von Proctoring bei (Online-) Tests. Anschließend werden für zwei ausgewählte dgp-Tests die durchschnittlichen Leistungen in beaufsichtigten und unbeaufsichtigten Online-Situationen miteinander verglichen und daraus Schlussfolgerungen gezogen.

Bisherige empirische Befundlage

In einer Übersichtsarbeit, einer sogenannten Metaanalyse, zu der Fragestellung, ob es einen Unterschied in den durchschnittlichen Testleistungen zwischen beaufsichtigten und unbeaufsichtigten Tests gibt, berücksichtigten Steger et al. (2020) insgesamt 49 Studien mit insgesamt N = 100.434 Testeilnehmer*innen. Es zeigte sich, dass unbeaufsichtigte Tests mit einem um 0.20 Standardabweichungen (STD) leicht höheren Test-Score einhergehen. **Im Schnitt schneiden die unbeaufsichtigten Personen besser ab als beaufsichtigte Personen.**

Die Autoren untersuchten mehrere Einflussfaktoren (Moderatoren) für diesen Effekt. Vor allem zwei Faktoren erwiesen sich hier als relevant:

- → Ob Test-Fragen durch eine einfache Internet-Recherche zu beantworten sind, hatte einen signifikanten Einfluss auf die Leistung in diesen Aufgaben bei unbeaufsichtigten Online-Tests. Wenn die richtigen Lösungen über das Internet hingegen nicht leicht zu finden waren, gab es fast keinen Unterschied in den Test-Scores zwischen unbeaufsichtigten und beaufsichtigten Online-Tests.
- → Für die Situation „high-stake“ (Testergebnis ist für Person bedeutsam, z. B. bei Testungen im Zuge von Personalauswahlverfahren) versus „low-stake“ (Testergebnis ist für Person nicht wichtig) konnte kein signifikanter Effekt gefunden werden. Auf deskriptiv-statistischer Ebene zeigte sich für high-stake Situationen aber ein Unterschied zwischen beaufsichtigten und unbeaufsichtigten Online-Tests von 0.27 STD. Für low-stake Situationen war der Effekt nur 0.09 STD. **Das bedeutet, dass in den vorliegenden Studien in unbeaufsichtigten Testungen eine erkennbar bessere Testleistung gezeigt wurde im Vergleich zu beaufsichtigten Testungen, wenn das Ergebnis des Tests für die Personen sehr relevant war.**

In einer Unterstichprobe von fünf Studien untersuchten Steger et al. (2020) die Stabilität der Rangreihenfolge der Ergebnisse von Testteilnehmer*innen unter Beaufsichtigung im Vergleich zu Nicht-Beaufsichtigung. Es zeigte sich ein moderater Zusammenhang der Testergebnisse mit und ohne Beaufsichtigung mit einer gepoolten Korrelation von =.58 (SE =.10), 95% CI [0.38, 0.78]. Die Autoren interpretieren dies als substanzielle Änderung der Rangreihenfolge durch die unterschiedlichen Testbedingungen. **Das bedeutet, dass Bewerber*innen unterschiedlich gute Plat-**

zierungen in der Bestenliste haben, je nachdem, ob beaufsichtigt wurde oder nicht.

In einer Studie von Hylton et al. (2016) mit webcambasierter Beaufsichtigung von Online-Tests wurden Unterschiede in der Bearbeitungszeit zwischen beaufsichtigten und nicht beaufsichtigten Online-Tests gefunden. Außerdem zeigte sich in dieser Studie ebenfalls, dass unbeaufsichtigte Tests im Durchschnitt zu höheren Test-Scores führten. Der Effekt fiel ähnlich hoch aus wie in der Metaanalyse von Steger et al. (2020). Zusätzlich berichteten Hylton et al. (2016), dass Testteilnehmer*innen in der nicht beaufsichtigten Testsituation eine größere Möglichkeit der Zusammenarbeit mit anderen wahrgenommen sowie eine größere Chance gesehen haben, nicht autorisierte Hilfsmittel zu benutzen.

Zusammenfassend zeigen die bisherigen empirischen Studien, dass unbeaufsichtigte Online-Tests im Schnitt zu besseren Ergebnissen führen als beaufsichtigte Online- oder Vor-Ort Tests. Auch für die dgp liegen mittlerweile erste Daten zum Vergleich von beaufsichtigten und unbeaufsichtigten Online-Tests vor.

Fallstudien: Unterschied zwischen beaufsichtigten und unbeaufsichtigten Online-Tests in der dgp

Die dgp führt seit Anfang 2023 beaufsichtigte Online-Tests inklusive Identifikation mit einem Ausweis durch. Das Proctoring-Angebot (weitere Informationen dazu unter dgp.de/proctoring) basiert auf der Beaufsichtigung per Kamera, Mikrofon sowie Screensharing. Die Testteilnehmer*innen müssen für die Beaufsichtigung keine Software installieren; es wird keine Kontrolle über den Rechner erzwungen. Die Daten werden auf Servern in Deutschland gespeichert, ausschließlich für den Zweck der Identitätsprüfung sowie Testbeaufsichtigung verwendet und anschließend gelöscht.

Der modulare Online-Test für Verwaltungen (MOT-V) der dgp wurde im beaufsichtigten Rahmen bereits mehrfach eingesetzt. Ebenfalls wurde der modulare Online-Test für gewerbliche Berufe (MOT-G) der dgp bei einer Behörde für identische Stellenausschreibungen in den Jahren 2022 online unbeaufsichtigt sowie im Jahr 2023 online beaufsichtigt durchgeführt. Für beide Online-Tests wird im Folgenden ein Vergleich der Ergebnisse von beaufsichtigten und unbeaufsichtigten Testungen vorgenommen.

Fragestellung und Hypothesen

Untersucht werden soll die Fragestellung, ob sich die Testergebnisse im Mittel bei beaufsichtigten und nicht beaufsichtigten Online-Tests in einer high-stake Situation unterscheiden. Auf Grund bisheriger empirischer Befunde ist zu erwarten, dass ein Unterschied vorliegt.

Fallstudie 1: Modularer Online-Test für Verwaltungsberufe

Methoden

Stichprobe

Eine beaufsichtigte Online-Testung haben N = 286 Bewerber*innen für die Laufbahn des mittleren öffentlichen Dienstes in Deutschland absolviert. Das durchschnittliche Alter lag bei M = 28.34 (SD = 9.58) Jahren. 51.05 % waren weiblich und 0.70 % divers. Der Zeitraum der Testdurchführung lag zwischen Februar und Mai 2023.

Eine unbeaufsichtigte Online-Testung haben N = 1439 Bewerber*innen für die Laufbahn des mittleren öffentlichen Dienstes in Deutschland absolviert. Das durchschnittli-

che Alter lag bei M = 26.93 (SD = 11.10) Jahren. 59.35 % waren weiblich und 0.90 % divers. Der Zeitraum der Testdurchführung lag zwischen September 2022 und Mai 2023.

Alle Bewerber*innen haben sich für den mittleren öffentlichen Dienst in Deutschland beworben. Die Qualifikation der Bewerber*innen ist somit vergleichbar, wobei keine weiteren Informationen zum Bildungsabschluss erfasst werden konnten. Jedoch sind die Stichproben keine Zufallsstichproben: Unter anderem können Selbstselektionsprozesse zu systematischen Unterschieden zwischen den Stichproben führen, weswegen aus dem Vergleich der Stichproben keine kausalen Schlussfolgerungen gezogen werden können. Dieser Umstand wird im Abschnitt der Diskussion erneut thematisiert.

Die Testdurchführung fand online über die Testplattform der dgp statt. Der Zugang zum Test erfolgt mittels einer individuellen, personenspezifischen TAN (Transaktionsnummer). Für die Testungen der dgp werden allgemein folgende Kontrollmechanismen abgesehen vom Proctoring verwendet: Zustimmung zu einem Vertrag, unbekannte Zeitbegrenzung der Aufgaben sowie die Ankündigung, dass Log-Daten der Bewerber*innen (z. B. beim Verlassen des Tabs) gespeichert werden. Letzteres wird nicht systematisch zum Ausschluss von Bewerber*innen verwendet.

Modularer Online-Test MOT-V

Alle Bewerber*innen absolvierten den MOT-V mit den Basismodulen verbale und numerische Verarbeitungskapazität auf Basis des Berliner Intelligenzstrukturmodells (Jäger et al., 1997). Nähere Informationen zur psychometrischen Qualität der Module sind unter dgp.de/eignungstests zu finden.

In der Stichprobe der beaufsichtigten Online-Testung haben sämtliche Bewerber*innen darüber hinaus alle in Tabelle 1 aufgeführten optionalen Module bearbeitet. In der Stichprobe der unbeaufsichtigten Testung haben nur Teile der Bewerber*innen die optionalen Module bearbeitet. Die Stichprobengrößen sind untenstehend in Tabelle 1 aufgeführt.

Tabelle 1: Übersicht der Stichprobengröße pro Testmodul

	Unbeaufsichtigt (N)	Beaufsichtigt (N)
Basismodule		
Verbale Verarbeitungskapazität	1439	286
NumerischeVerarbeitungskapazität	1439	286
Optionale Module		
Arbeitseffizienz	1412	286
Rechtschreibung	1412	286
Allgemeinwissen	1182	286
Verwaltungswissen	539	286
Wirtschaftswissen	390	286

Geplante Auswertung

Auf Grund der nicht völlig vergleichbaren Stichproben und der unterschiedlichen Stichprobengrößen werden nicht die direkten mittleren Leistungen der Bewerber*innen in beaufsichtigten und unbeaufsichtigten Online-Tests verglichen. Stattdessen werden aus beiden Gruppen 1000 Zufallsstichproben von N = 100 Bewerber*innen gezogen, wobei pro Geschlecht kontrolliert wird, indem immer je 50 Männer und Frauen gezogen werden.

Die Unterschiede in den Ergebnissen zwischen beaufsichtigten und unbeaufsichtigten Online-Tests werden für alle 13 Testbereiche als über 1000 Zufallsstichproben gemittelte standardisierte Mittelwertsunterschiede (Cohens d) berichtet. Für jede Zufallsstichprobe wird ein t-Test für unabhängige Stichproben durchgeführt und gezählt, wie häufig der p-Wert (Signifikanzwert bzw. Überschreitungwahrscheinlichkeit) des t-Tests unterhalb des alpha-Niveaus von 0.05 liegt.

Ergebnisse

Die Ergebnisse in Tabelle 2 zeigen, dass bei allen Testbereichen/Modulen die Anzahl korrekt gelöster Aufgaben im unbeaufsichtigten Test höher war im Vergleich zum beaufsichtigten Test. Der Ausmaß der Differenz variiert mit kleinen bis moderaten Effekten zwischen Cohens d = .101 bis zu Cohens d = .626. Weitgehend zeigen sich sehr kleine Unterschiede. Bei den Testteilen Grundrechnen sowie Allgemeinwissen sind die Effekte deutlich höher.

Tabelle 2: Mittelwertsunterschiede (Cohens d) sowie Anzahl signifikanter t-Tests

Testbereich	Cohens d	Häufigkeit p < 0.05
Textrechnen	.361	751
Zahlenreihen	.279	500
Grundrechnen	.626	996
Tabellen und Statistiken	.131	74
Analogien	.108	38
Textanalyse	.101	31
Wortklassifikationen	.109	42
Schlüsse vergleichen	.135	90
Reisekosten	.103	27
Rechtschreibung	.231	340
Allgemeinwissen	.611	998
Verwaltungswissen	.115	48
Wirtschaftswissen	.112	47

Fallstudie 2: Modularer Online-Test für gewerbliche Berufe

Methoden

Stichprobe

Im Jahr 2022 haben N = 135 Bewerber*innen den modularen Online-Test für gewerbliche Berufe unbeaufsichtigt absolviert. Davon waren 23 Personen weiblich. Das mittlere Alter betrug M = 23.36 Jahre.

Im Jahr 2023 haben N = 145 Bewerber*innen den modularen Online-Test für gewerbliche Berufe beaufsichtigt absolviert. Davon waren 29 Personen weiblich. Das mittlere Alter betrug M = 21.58 Jahre.

Die Bewerber*innen haben sich in beiden Jahren auf die gleiche Stellenausschreibung bei der gleichen Behörde beworben.

Die Testdurchführung fand online über die Testplattform der dgp statt. Der Zugang zum Test erfolgt mit einer individuellen, personenspezifischen TAN. Mit Blick auf den Testzugang sowie auf die angewandten allgemeinen Kontrollmechanismen gelten die zur Fallstudie 1 gegebenen Hinweise analog.

Modularer Online-Test MOT-G

Alle Bewerber*innen absolvierten den MOT-G mit den Basismodulen verbale, numerische und figurale Verarbeitungskapazität sowie Bearbeitungsgeschwindigkeit auf Basis des Berliner Intelligenzstrukturmodell (BIS). Dazu wurden noch die Module Praktisches Verständnis, Deutsch Rechtschreibung sowie Deutsch Grammatik durchgeführt. Nähere Informationen zur psychometrischen Qualität der Module sind unter dgp.de/eignungstests zu finden.

Geplante Auswertung

Die Ergebnisse der Bewerber*innengruppe liegen in Standardwerten (M = 100, SD = 10) vor. Für jedes Modul werden die Testergeb-

Tabelle 3: Ergebnisse der t-Tests

Modul	M unbeaufsichtigt	M beaufsichtigt	Cohens d	t	df (Freiheitsgrad)	p-Wert (Wert der Teststatistik)
Verbale Verarbeitungskapazität	103.02	98.52	.33	2.76	277.07	.01
Numerische Verarbeitungskapazität	102.47	98.52	.43	3.64	277.88	.00
Figurale Verarbeitungskapazität	104.30	102.32	.23	1.87	271.41	.06
Numerische Bearbeitungsgeschwindigkeit	102.57	98.79	.31	2.56	277.96	.01
Figurale Bearbeitungsgeschwindigkeit	96.18	95.26	.10	.84	277.67	.40
Praktisches Verständnis	104.96	103.37	.19	1.58	277.96	.11
Deutsch Rechtschreibung	101.20	95.01	.75	6.26	274.27	.00
Deutsch Grammatik	100.87	99.12	.21	1.77	276.39	.08
Gesamttestwert	**101.93**	**93.30**	**.43**	**3.56**	**227.91**	**0.00**

nisse des unbeaufsichtigten und beaufsichtigten Testdurchlaufs mit Hilfe der Mittelwerte sowie der standardisierten Mittelwertsdifferenz Cohens d gegenübergestellt. Der Unterschied der Testergebnisse wird je mit einem t-Test inferenzstatistisch abgesichert.

Ergebnisse

Die Ergebnisse zeigen bei allen Modulen sowie beim Gesamttest, dass die Testergebnisse im unbeaufsichtigten Test leicht höher lagen. Die Testergebnisse von unbeaufsichtigten und beaufsichtigten Testungen unterscheiden sich bei vier Modulen signifikant voneinander: Verbale und numerische Verarbeitungskapazität, numerische Bearbeitungsgeschwindigkeit sowie Rechtschreibung haben in der unbeaufsichtigten Testung signifikant höhere Werte. Auch für den Gesamttestwert zeigt sich ein signifikant höheres Ergebnis für den unbeaufsichtigten Test im Jahr 2022. Die standardisierten Mittelwertdifferenzen bewegen sich mit Ergebnissen zwischen Cohens d = .10 und Cohens d = .75 im Bereich kleiner bis mittlerer Effekte.

Diskussion & Ausblick

Sowohl die bisherigen Studien als auch die dgp-Fallstudien zeigen übereinstimmend, dass bei unbeaufsichtigten Online-Tests im Mittel bessere Ergebnisse erzielt werden. Das ist ein deutlicher Hinweis darauf, dass – trotz Gegenmaßnahmen wie Zeitbegrenzungen und Zustimmung zu einem Vertrag – bei Online-Tests in high-stake Situationen erfolgreiche Täuschungen unternommen werden.

Einschränkungen

Es ist zunächst nur eine Vermutung, dass die beobachteten Mittelwertsunterschiede zwischen beaufsichtigten und unbeaufsichtigten Online-Tests durch Betrugsversuche zustande kommen. Eine empirische Untersuchung von Betrugsversuchen durch Testteilnehmer*innen ist insbesondere in high-stake Situationen nur schwer möglich. Um sicher sein zu können, wie sich Betrugsversuche von Testteilnehmer*innen auswirken, müsste man dieses experimentell untersuchen, d. h. Gruppen, die betrügen, mit Gruppen, die nicht betrügen, vergleichen. In der Praxis sind bewusste Einflussnahmen in einer Personalauswahlsituation jedoch ethisch und rechtlich nicht vertretbar. Daher beziehen wir uns auf die Daten, die bereits vorhanden sind (Beobachtungsstudien), und ziehen aus den Beobachtungen vorläufige Schlussfolgerungen.

Mit Umsicht aus den vorhandenen Unterschieden Schlussfolgerungen zu ziehen, bedeutet auch, sich zu fragen, woher die Unterschiede ansonsten stammen könnten. Welche alternativen Erklärungen gibt es für Mittelwertsunterschiede zwischen beaufsichtigten und unbeaufsichtigten Testungen, die nicht auf Betrug zurückzuführen sind?

Naheliegend für die Daten des MOT-V sind in der vorliegenden Fallstudie systematische Unterschiede zwischen den Stichproben. Die Stichproben sind keine Zufallsstichproben, sondern stammen von unterschiedlichen öffentlichen Behörden, bei denen sich Personen möglicherweise mit systematisch unterschiedlicher Motivation, Persönlichkeit oder Leistungsfähigkeit beworben haben. Wir haben in diesen Fallstudien versucht, dieser Problematik zu begegnen, indem wir aus den vorhandenen Gruppen zufällige Stichproben

gezogen haben. Dennoch lässt sich damit ein systematischer Unterschied zwischen den Gruppen aufgrund möglicherweise anderer Variablen nicht ganz ausschließen.

Die Schlussfolgerungen aus den MOT-V Daten sind in dieser Hinsicht eingeschränkt. Sie sind dennoch wertvoll, um einen Einblick in mögliche Unterschiede zwischen beaufsichtigten und unbeaufsichtigten Testungen zu gewinnen. Ergänzend dazu ermöglicht die Fallstudie mit Daten aus dem gewerblichen Bereich, die Unterschiede zwischen beaufsichtigten und unbeaufsichtigten Testungen für eindeutig vergleichbare Bewerber*innengruppen bei der gleichen Behörde zu beurteilen. Systematische Unterschiede zwischen den Bewerber*innengruppen sind hier nicht auf Grundlage der Wahl der Behörde für die Bewerbung zu erwarten. Eine mögliche alternative Erklärung für Unterschiede könnte in dieser Stichprobe lediglich der Zeitraum zwischen den Testungen sein.

Eine weitere alternative Erklärung für Unterschiede zwischen beaufsichtigten und unbeaufsichtigten Testungen ist Testangst. Studien deuten darauf hin, dass bei beaufsichtigten Online-Tests die Testangst höher ist im Vergleich zu unbeaufsichtigten Online-Tests (Karim et al., 2014; Lilley et al., 2016; Stowell & Bennett, 2010). Vermutlich ist auch die Testangst bei beaufsichtigen Präsenz-Tests höher als bei unbeaufsichtigten Online-Tests. Erhöhte Testangst hat möglicherweise Auswirkungen auf die Testleistung.

Die in der Literatur berichteten sowie in den Fallstudien beobachteten Effekte sind klein, aber über viele Studien stabil. In der wissenschaftlichen Literatur wird „klein“ manches Mal gleichgesetzt mit „unbedeutsam“. Aus praktischer Perspektive ergibt sich aus unserer Sicht allerdings ein anderes Bild auf diesen Sachverhalt. Es ist nicht davon auszugehen, dass bei unbeaufsichtigten Online-Tests alle Personen gleichermaßen betrügen, sondern dass es nur einen kleineren Anteil von Täuschungen gibt. Diese Täuschungshandlungen sind aller Wahrscheinlichkeit nach von weiteren Eigenschaften und situativen Faktoren abhängig, beispielweise Persönlichkeit, Motivation o. ä.. Wir erwarten deswegen keine großen Effekte auf Gruppenebene, sondern ebendiese beobachteten kleinen Unterschiede.

Für die Praxis relevant ist jedoch nicht die Ebene eines Gruppenvergleichs, sondern die der Einzelentscheidungen. Ein minimaler Unterschied auf Gruppenebene bedeutet in diesem Fall also wahrscheinlich nicht, dass alle Bewerber*innen in unbeaufsichtigten Online-Tests leicht besser sind, weil sie betrügen, sondern dass einzelne Bewerber*innen deutlich besser sind, weil sie betrügen. Diese Bewerber*innen haben dann in einer Auswahlsituation im Rangreihenfolgenvergleich eine deutlich erhöhte Chance, einen Schritt im Auswahlverfahren weiterzukommen. Auch empirisch zeigte sich bei Steger et al. (2020), dass die Rangreihenfolge sich bei beaufsichtigten und unbeaufsichtigten Online-Tests verändert.

Aus praktischer Sicht ist das auch deswegen hoch relevant, weil dadurch eine deutliche Ungleichbehandlung von Bewerber*innen entsteht, eine „unfaire“ Situation, in der Bewerber*innen immer davon ausgehen müssen, dass andere betrügen und sich dadurch Vorteile in der Rangreihenfolge verschaffen. Diese Situation könnte wiederum dazu führen, dass die Motivation für Täuschungsversuche insgesamt erhöht wird.

Ausblick

Beaufsichtigte Online-Tests sind erst seit Beginn 2023 in der dgp möglich. Mit der Zeit werden wir mehr und mehr Daten sammeln, die Auskunft über die Unterschiede zwischen beaufsichtigten und unbeaufsichtigten Online-Tests geben werden. Dann sind weitere Untersuchungen möglich.

Im Kontext der Personalauswahl ist zusätzlich die Frage relevant, ob die prädiktive Validität von Tests bezüglich der Arbeitsleistung dadurch beeinflusst wird, dass diese als unbeaufsichtigte Online-Tests durchgeführt werden. Eine Studie von Beaty et al. (2011) hat gezeigt, dass die Kriteriumsvalidität von beaufsichtigten und unbeaufsichtigten Tests bezüglich Arbeitsleistung vergleichbar bleibt. Weitere Untersuchungen stehen hier aus.

Des Weiteren ist interessant, ob die Beaufsichtigung von Online-Tests das gemessene Konstrukt kognitiver Fähigkeiten verändert (Steger et al., 2020). Diese Fragestellung ist in der Wissenschaft unter dem Begriff „Messinvarianz der Tests“ bekannt. Um beaufsichtigte und unbeaufsichtigte Online-Tests miteinander vergleichen zu können, sollte die Messung über beide Situationen hinweg invariant sein, d. h. von der Struktur her identisch.

Fazit

Die vorliegenden Daten sowie Ergebnisse von wissenschaftlichen Studien zeigen eindeutig, dass die Testleistung bei unbeaufsichtigten Tests im Mittel höher ausfällt als bei beaufsichtigten. Das deutet darauf hin, dass bei unbeaufsichtigten Online-Tests von Bewerber*innen getäuscht wird. Täuschungen bei Online-Tests sind nicht ein Problem auf Ebene von Gruppen (nicht alle Bewerber*innen betrügen!), sondern für konkrete Einzelfälle, die sich damit einen möglicherweise deutlichen Vorteil im Auswahlverfahren verschaffen können. Die Möglichkeit der Täuschung bei unbeaufsichtigten Online-Tests gefährdet damit die Schlussfolgerungen, die aus Online-Tests gezogen werden können. Da andere Gegenmaßnahmen allein nicht wirksam sind, ist die Beaufsichtigung von Online-Test das Mittel der Wahl, um die Vorteile von Online-Tests zu nutzen, und um gleichzeitig die Nachteile zu verringern, indem allen Bewerber*innen ein fairer Auswahlprozess ermöglicht wird.

LITERATUR

Beaty, J. H., Nye, C. D., Borneman, M. J., Kantrowitz, T. M., Drasgow, F. & Grauer, E. (2011). Proctored Versus Unproctored Internet Tests: Are unproctored noncognitive tests as predictive of job performance? International Journal of Selection and Assessment, 19(1), 1–10. https://doi.org/10.1111/j.1468-2389.2011.00529.x

Hylton, K., Levy, Y. & Dringus, L. P. (2016). Utilizing webcam-based proctoring to deter misconduct in online exams. Computers & education, 92–93, 53–63. https://doi.org/10.1016/j.compedu.2015.10.002

Jäger, A. O., Süß, H. & Beauducel, A. (1997). Berliner Intelligenzstruktur-Test : BIS-Test. Hogrefe. https://madoc.bib.uni-mannheim.de/14578/

Jobmann, A. & Kleinmanns, A. (2023). Beaufsichtigung von Online-Tests: Ja oder nein? dgp informationen 2023/24

Karim, M. N., Kaminsky, S. E. & Behrend, T. S. (2014). Cheating, Reactions, and Performance in Remotely Proctored Testing: An Exploratory Experimental Study. Journal of Business and Psychology, 29(4), 555–572. https://doi.org/10.1007/s10869-014-9343-z

Lilley, M., Meere, J. & Barker, T. (2016). Remote Live Invigilation: A Pilot Study. Journal of interactive media in education, 2016(1). https://doi.org/10.5334/jime.408

Stowell, J. R. & Bennett, D. (2010). Effects of Online Testing on Student Exam Performance and Test Anxiety. Journal of Educational Computing Research, 42(2), 161–171. https://doi.org/10.2190/ec.42.2.b

Steger, D., Wilhelm, O. & Gnambs, T. (2020). A Meta-Analysis of Test Scores in Proctored and Unproctored Ability Assessments. European Journal of Psychological Assessment, 36(1), 174–184. https://doi.org/10.1027/1015-5759/a000494

Kontakt

Dr. Anna-Lena Jobmann
Diplom-Psychologin
Deutsche Gesellschaft für
Personalwesen e. V.

Dr. Anna-Lena Jobmann ist Mitarbeiterin der Stabsstelle Forschung und Entwicklung der dgp e. V. Ihre Schwerpunkte sind die Entwicklung und Überprüfung von kognitiven Eignungstests auf Basis bewährter statistischer Testmodelle, inklusive adaptiven Testens, Messung sozialer Kompetenzen und Fähigkeiten sowie Fairness von Eignungstests.

jobmann@dgp.de

Amelie Kleinmanns
M. Sc. Psychologie
Deutsche Gesellschaft für
Personalwesen e. V.

Amelie Kleinmanns ist Mitarbeiterin der Stabsstelle Forschung und Entwicklung der dgp e. V. Ihre Schwerpunkte sind die Entwicklung und Überprüfung von kognitiven Eignungstests sowie die Entwicklung von Persönlichkeitsinventaren auf Basis bewährter statistischer Testmodelle.

kleinmanns@dgp.de