

PERSEO Validierungsstudie 2023

Messung der Vorhersagekraft unserer Tests

Um die Vorhersagekraft der PERSEO Leistungstests zu überprüfen, haben wir eine empirische Untersuchung mit 99 Teilnehmern durchgeführt. Hierbei wurde der statistische Zusammenhang zwischen dem PERSEO Testergebnis und der Leistung in Ausbildung und Studium ermittelt. Der unkorrigierte Validitätskoeffizient beträgt $r = .48$; korrigiert um Varianzeinschränkung: $r = .56$. Die Ergebnisse zeigen, dass PERSEO Leistungstests Ausbildungs- und Studienleistungen zuverlässig vorhersagen.

Die Validität ist neben der Objektivität und der Reliabilität eines der drei Hauptgütekriterien psychologischer Tests. Sie verrät uns, ob der Test das misst, was er messen soll. Bei beruflichen Eignungstests ist vor allem die prognostische Validität wichtig. Sie besagt, wie gut das Testergebnis die spätere Leistung der Kandidaten vorhersagt.

Eignungstests sollten also immer eine möglichst hohe Validität besitzen: Teilnehmer, die in einem validen Test gut abschneiden, haben im Schnitt bessere Noten und fallen seltener durch Prüfungen. Eine hohe Validität hat für den auswählenden Arbeitgeber auch einen finanziellen Wert – er spiegelt sich etwa in einem geringeren Betreuungsaufwand, einem niedrigeren Abbruchrisiko von Ausbildung oder Studium und einem geringeren Verlustrisiko der damit verbundenen zeitlichen und monetären Investitionen wider.

Mathematisch ist die prognostische Validität ein Wert zwischen 0 und 1, der *Validitätskoeffizient* genannt wird.

Dr. Benjamin Haarhaus, Dipl.-Psych.
PERSEO GmbH, Graf-Adolf-Str. 41, 40210 Düsseldorf
info@perseo.hr, www.perseo.hr

Je größer der Wert, umso zuverlässiger die Vorhersage. Aber wie groß sollte der Wert sein, damit man von einem „guten Test“ sprechen kann? In einer Meta-Analyse (Hülshager et al., 2006) wurde ermittelt, dass der (unkorrigierte) Zusammenhang von Intelligenztests und Ausbildungserfolg in Deutschland bei $r = .32$ liegt, der korrigierte Wert beträgt $r = .47$ (was es mit der Korrektur auf sich hat, erläutern wir weiter unten).

Rechtssicherheit und Qualitätsvorgaben

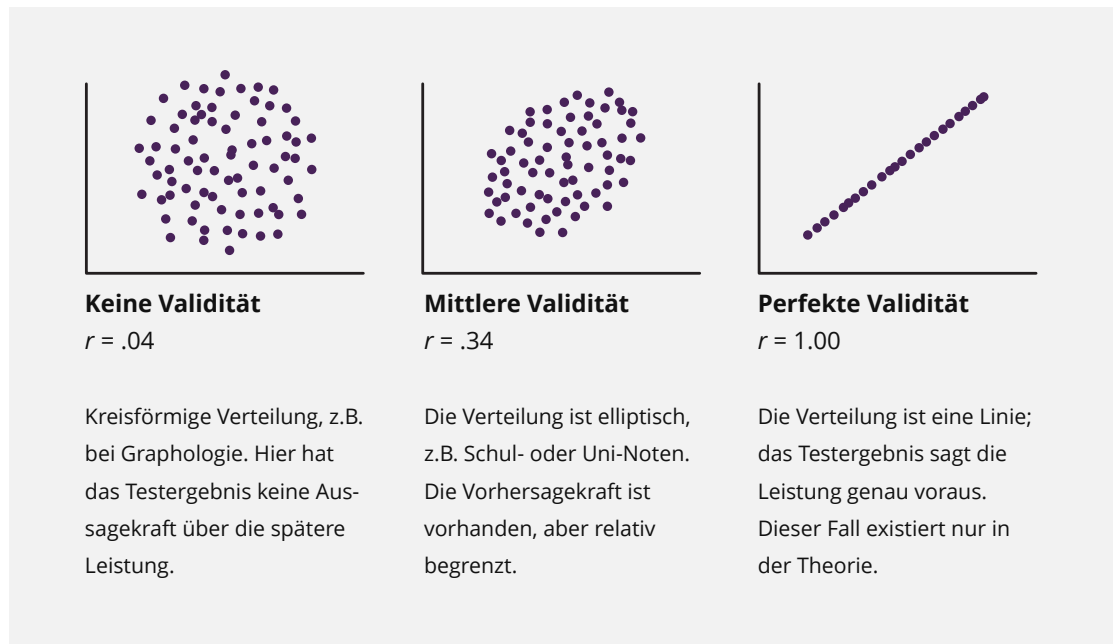
Als Testanbieter legen wir Wert auf Wissenschaftlichkeit – denn nur durch sorgfältig konstruierte und evaluierte Verfahren können wir unseren Kunden gute und rechtssichere Auswahlentscheidungen ermöglichen. Eine wichtige Voraussetzung dafür ist die Validität unserer Tests, die wir hiermit empirisch nachweisen. Die vorliegende Studie haben wir mit Unterstützung unserer Kunden durchgeführt. Mit ihr möchten wir sowohl den Vorgaben der DIN 33430 nachkommen als auch die Qualität unserer Verfahren evaluieren und offenlegen. Die Rohdaten der Untersuchung stellen wir Ihnen gerne zur Verfügung.

ABBILDUNG 1

Verschiedene Validitäten, schematische Darstellung

ONLINE-RECHNER

Welche Auswirkung die Validität auf die Trefferquote hat, können Sie hier ausprobieren: www.perseo.hr/de/trefferquote



Empirische Untersuchung

Vorgehen

Die Datenerhebung erfolgte von August bis November 2022. Verschiedene Auftraggeber wurden über die geplante Studie informiert und zur Teilnahme eingeladen. Hierfür baten wir darum, uns die Teilnehmernummern der eingestellten Bewerber sowie ein Leistungsmaß (Noten) aus Studium oder Ausbildung zu übermitteln. Über die Teilnehmernummern konnten wir die erzielten Testergebnisse ermitteln und mit den Leistungsmaßen in Beziehung setzen.

Stichprobe

Insgesamt wurden die Daten von 99 Teilnehmern übermittelt. Die Teilnehmer (62 % weiblich) waren zum Zeitpunkt der Testdurchführung durchschnittlich 20,2 (Standardabweichung: 4,4) Jahre alt. Es handelt sich um Personen, die sich auf eine Ausbildung (80 %) oder ein duales Studium (20 %) im öffentlichen Sektor beworben haben. Der Großteil (72 %) entfällt auf den Verwaltungsbereich (Verwaltungsfachangestellte, Bachelor-Studium Öffentliche Verwaltung, Fachangestellte für Medien- und Informationsdienste), 19 % auf den

gewerblich-technischen Bereich (Fachkraft für Kreislauf- und Abfallwirtschaft, Gärtner, KFZ-Mechatroniker, Straßenbauer) und 9 % auf den IT-Bereich (Fachinformatiker). 61 % der Teilnehmer haben den Test unter Aufsicht absolviert, 39 % als unbeaufsichtigten Onlinetest von Zuhause.

Die Daten stammen von fünf Auftraggebern (AG) aus dem öffentlichen Sektor. Bei AG 1 und 2 handelt es sich um Großstädte mit 160.000 und 110.000 Einwohnern. AG 3, 4 und 5 sind Kreisverwaltungen von Landkreisen mit je über 200.000 Einwohnern.

TABELLE 1

Deskriptive Statistiken und Interkorrelationen

	<i>M</i>	<i>SD</i>	1	2	3
1. Alter	20.15	4.39	-	.34	.16
2. Gesamtpunktwert	61.08	13.17		-	.24
3. Note	10.04	2.96			-

Hinweis: *M* = Mittelwert, *SD* = Standardabweichung, 1, 2, 3 = Interkorrelationen

Auswertung

Um die Vorhersagekraft eines Tests zu ermitteln, werden zwei Werte statistisch in Beziehung gesetzt: Eine vorhersagende Variable (Prädiktor) und eine vorhergesagte (Kriterium).

Prädiktor: Diese Variable ist der Gesamtpunkt看wert im PERSEO Test. Dieser ergibt sich als gewichteter Mittelwert aus den einzelnen Aufgabenbereichen. Die meisten PERSEO Tests für Azubis und duale Studenten enthalten Aufgaben aus den Bereichen Logik, Konzentration, Deutsch und Mathematik. Da sich die Teilnehmer bei unterschiedlichen Behörden und für unterschiedliche Laufbahnen und Fachrichtungen beworben haben, sind die durchgeführten Tests jedoch zum Teil unterschiedlich. So enthalten z.B. die Tests für Verwaltungsberufliche Aufgaben zum Verständnis von Gesetzestexten, gewerbliche Ausbildungen hingegen Aufgaben zum praktisch-technischen Verständnis.

Kriterium: Die Variable, die vorhergesagt werden soll, ist die Leistung in Ausbildung und Studium. Unsere Auftraggeber haben uns verschiedene Leistungsmaße übermittelt: Die Durchschnittsnote nach dem ersten Berufsschuljahr bzw. zwei Semestern (56 %), die Durchschnittsnote nach dem zweiten Berufsschuljahr bzw. vier Semestern (11 %), die Note der Zwischenprüfung (26%) und die Durchschnittsnote des Abschlusszeugnisses (7 %).

Um die Lesbarkeit der Ergebnisse zu verbessern, haben wir die Noten von einer 6er- in eine 15er-Skala überführt (15 Punkte: 1+, 14 Punkte: 1, 13 Punkte: 1- usw.), da bessere Leistungen so mit größeren Zahlenwerten einhergehen.

Um den Zusammenhang zwischen dem PERSEO Gesamtpunkt看wert (GPW) und den Ausbildungs- bzw. Studienleistungen zu ermitteln, wurden bivariate Korrelationen genutzt. Da sich die durchschnittlichen Punktwerte und Noten zwischen den Auftraggebern z.T. enorm unterscheiden, ist die Vergleichbarkeit aller-

dings stark eingeschränkt: So fallen z.B. die übermittelten Zwischenprüfungsnoten deutlich schlechter aus ($M = 8.3$) als die der Abschlussprüfungen ($M = 12.0$). Teilnehmer, die den Test ohne Aufsicht von Zuhause aus durchgeführt haben, erzielten durchschnittlich bessere Testergebnisse als die, die den Test unter Aufsicht vor Ort absolviert haben (vgl. Tabelle 2). Diese Kombination verschiedener Durchführungs- und Leistungsdaten führt statistisch zu einem niedrigen mittleren Zusammenhang ($r = .24$), obwohl der Zusammenhang bei allen Auftraggebern eigentlich höher ist (AG 1: .55, AG 2: .63, AG 3: .33, AG 4: .31, AG 5: .76).

TABELLE 2

Statistiken nach Durchführungsart und Art der Leistungsdaten

Gruppe		n	GPW	Note
Durchführung	ohne Aufsicht	39	66.13	9.53
	mit Aufsicht	60	57.82	10.37
Leistungsdaten	Durchschnitt Jahr 1	55	61.85	10.50
	Durchschnitt Jahr 2	11	60.82	10.48
	Zwischenprüfung	26	61.73	8.33
	Abschlusszeugnis	7	53.14	11.99

Hinweis: n = Stichprobenumfang, GPW = Gesamtpunkt看wert im PERSEO Test

Um Fehlinterpretationen zu vermeiden, wurden die Punktwerte und Noten je Auftraggeber zentriert. Dabei wird der Mittelwert der jeweiligen Variable von den einzelnen Werten dieser Variable abgezogen, so dass ein Mittelwert von Null resultiert. Dieses Vorgehen macht die Daten der verschiedenen Auftraggeber vergleichbar und ermöglicht die Berechnung eines aussagekräftigen Gesamtwertes.

TABELLE 3

Deskriptive Statistiken und Validitätskoeffizienten nach Durchführung, Fachrichtung und Laufbahn

Kategorie		<i>n</i>	GPW	Note	<i>r</i>	<i>r</i> _{korrigiert}
Durchführung	ohne Aufsicht	39	66.13 (09.48)	9.53 (3.34)	.36	.52
	mit Aufsicht	60	57.82 (14.23)	10.37 (2.68)	.54	.56
Fachrichtung	Verwaltung	71	63.42 (12.00)	10.26 (2.85)	.36	.43
	Gewerblich	19	54.68 (15.98)	10.01 (3.22)	.80	.80
	Informatik	9	56.22 (10.89)	8.37 (3.11)	.26	.35
Laufbahn	Ausbildung	79	59.05 (13.17)	10.34 (2.96)	.49	.54
	Studium	20	69.15 (09.77)	8.85 (3.33)	.49	.65

Hinweis: *n* = Stichprobenumfang, GPW = Gesamtpunktwert im PERSEO Test, *r* = Validitätskoeffizient, Korrektur nach Thorndike Case 2, Standardabweichungen in Klammern

Ergebnisse

Der ermittelte Validitätskoeffizient für die Gesamtstichprobe beträgt $r = .48$. Im nächsten Schritt wurde dieser Wert um Varianzeinschränkung korrigiert (Schmidt, Shaffer, & Oh, 2008). Grund hierfür ist, dass in der Regel nur Bewerber mit guten und sehr guten Testergebnissen eingestellt werden – folglich fällt die Streuung der Testergebnisse (Varianz) unter den ausgewählten Bewerbern geringer aus als unter allen Testteilnehmern. Dies führt dazu, dass der statistische Zusammenhang zwischen Gesamtpunktwert und Note unterschätzt wird. Um diesen Effekt auszugleichen, ist es üblich, die Varianzeinschränkung mathematisch zu korrigieren. Hierfür haben wir die *Thorndike Case 2 Methode* genutzt (Thorndike, 1947). Der auf diese Weise korrigierte Koeffizient beträgt $r = .56$. Weitere Korrekturen, z.B. für Reliabilitätseinschränkungen, haben wir nicht vorgenommen.

Um detailliertere Einblicke zu erhalten, haben wir die Korrelationen außerdem nach Art der Durchführung, Fachrichtung und Laufbahn getrennt berechnet (vgl. Tabelle 3). Hierbei ist allerdings einschränkend zu beachten, dass die Stichprobenumfänge zum Teil recht gering sind. Die Ergebnisse haben daher eher explorativen Charakter.

Bei der **Art der Durchführung** fällt der ermittelte Zusammenhang zwischen Punktwert und Leistung für Vor-Ort-Testungen höher aus als für Onlinetestungen ohne Aufsicht. Dies ist allerdings durch die höhere Streuung der Vor-Ort-Testungen zu erklären. Nach Korrektur um Varianzeinschränkung sind beide Koeffizienten ähnlich hoch. Betrachtet man die verschiedenen **Fachrichtungen**, fällt der hohe Zusammenhang der gewerblichen Ausbildungen ins Auge. Auch hier hängt der starke Zusammenhang mit der großen Streuung der Testergebnisse zusammen. Praktisch bedeutet das: Es wurden auch Teilnehmer mit schlechten Testergebnissen eingestellt, die später auch entsprechend schwache Leistungen in der Ausbildung zeigten (vgl. Abbildung 2). Für die verschiedenen **Laufbahnen** (Ausbildung und Studium) ergeben sich vergleichbare Validitätskennwerte.

Diskussion

Die Ergebnisse zeigen, dass PERSEO Leistungstests spätere Ausbildungs- und Studienleistungen zuverlässig vorhersagen. Der ermittelte Validitätskoeffizient von $r = .48$ ist substantiell und mit den Ergebnissen früherer Untersuchungen (Hülshager et al., 2006) vergleichbar. Die Einzelergebnisse weisen zudem darauf

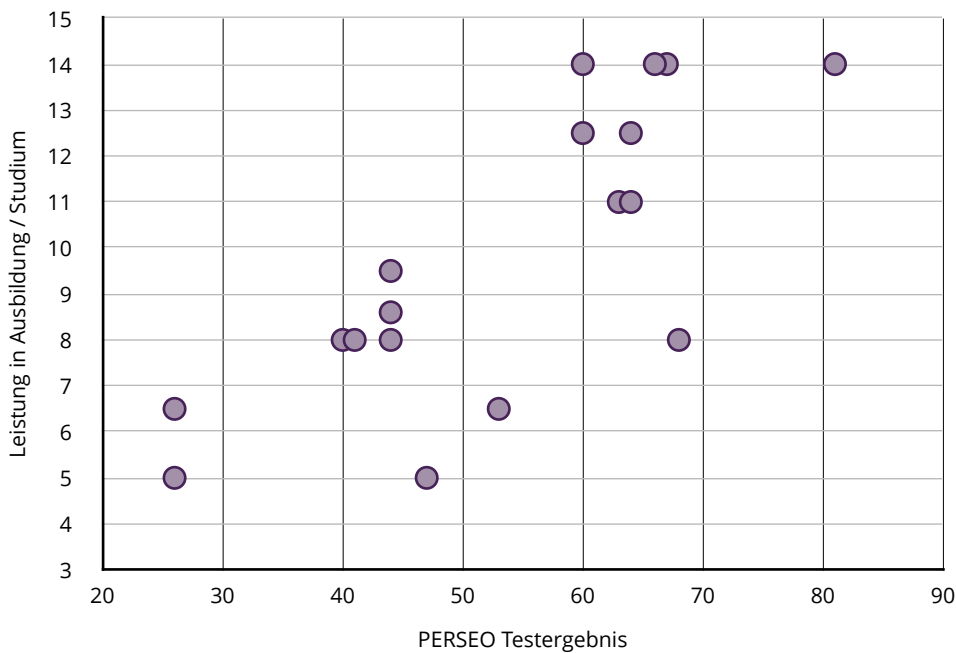


ABBILDUNG 2
Zusammenhang von PERSEO
Testergebnis und späterer
Leistung bei gewerblichen
Azubis (Daten von AG 1)

hin, dass die Vorhersagen bei verschiedenen Berufsgruppen, Durchführungsarten und Laufbahnen treffsicher sind.

Einschränkend ist anzumerken, dass der Stichprobenumfang unserer Untersuchung relativ niedrig ist. Das beeinflusst vor allem die Einzelkorrelationen, da diese anfälliger für Verzerrungen durch Ausreißer sind und größere Standardfehler haben. Wichtig ist außerdem, dass der ermittelte Validitätskoeffizient ein Durchschnittswert ist, der auf verschiedenen Tests für verschiedene Berufsgruppen basiert. Er spiegelt daher nicht eine „endgültige Validität“ wider, sondern kann je nach Testzusammenstellung und -gewichtung auch höher oder niedriger ausfallen. Aus diesem Grund werden wir in Zukunft weitere Validierungsstudien dieser Art durchführen, um die vorliegenden Ergebnisse zu replizieren und zu erweitern.

Literatur

- Hülshager, U. R., Maier, G. W., Stumpp, T., & Muck, P. M. (2006). Vergleich kriteriumsbezogener Validitäten verschiedener Intelligenztests zur Vorhersage von Ausbildungserfolg in Deutschland: Ergebnisse einer Metaanalyse. *Zeitschrift für Personalpsychologie*, 5(4), 145-162.
- Schmidt, F. L., Shaffer, J. A., & Oh, I. S. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology*, 61(4), 827-868.
- Thorndike, R. L. (1947). *Research problems and techniques* (Rep. No. 3 AAF Aviation Psychology Program Research Reports). Washington, DC: U.S. Government Printing Office.

Sie haben Interesse, an zukünftigen Studien mitzuwirken?

Wir würden uns freuen, Sie hierzu in unserem Kundenkreis begrüßen zu dürfen. Besuchen Sie uns jetzt auf www.perseo.hr und testen Sie uns kostenlos und unverbindlich.

Perseo GmbH

Rehgraben 9
47809 Krefeld
+49 2151 971 964-0

info@perseo.hr

www.perseo.hr