Testvorstellung der IRT-basierten dgp-Testbatterie VK-1

Anna-Lena Jobmann

Allgemeine Informationen

Der VK-1 ist ein computerbasierter Kurz-Test, der seit 2020 zur Messung von kognitiver Verarbeitungskapazität im Rahmen einer multimodalen Auswahlstrategie (z.B. im Assessment-Center) eingesetzt werden kann. Durch die kurze Gesamtdauer, kombiniert mit einer sehr hohen psychometrischen Qualität, liefert der Test eine zuverlässige und valide Einschätzung der grundlegenden kognitiven Fähigkeiten. Basierend auf dem Berliner Intelligenzstruktur-Modell, erfasst der Test die drei Inhaltsbereiche verbale, numerische und figurale Verarbeitungskapazität mit je einem Testbereich. Diese Testbereiche wurden auf Basis der Item Response Theorie konstruiert und sind rasch homogen. Der VK-1 kann als Einzel- oder Kleingruppentestung am Tablet oder Laptop sowie als unbeaufsichtigter Online-Test durchgeführt werden und dauert ca. 1 Stunde. Die Aufgabenbearbeitung ist zeitlich begrenzt.

Theoretische Grundlage als Ausgangspunkt der Testkonstruktion

Die theoretische Grundlage der Testbatterie ist das Berliner Intelligenzstruktur-Modell (Jäger, Süß & Beauducel, 1997). Kognitive Fähigkeiten werden in diesem Modell in Operationen und Inhalte gegliedert. Auf Seiten der Operationen erfasst die vorliegende Testbatterie die der kognitiven Verarbeitungskapazität, welche empirisch als die mit dem größten Vorhersageerfolg für berufliche Leistungen gilt (z.B. Ones, Viswesvaran & Dilchert, 2005). Die Verarbeitungskapazität wird in den Inhaltsbereichen figural, numerisch und verbal erfasst (Jäger, Süß & Beauducel, 1997). Die figurale Verarbeitungskapazität bezeichnet die Fähigkeit, komplexe bildhafte Informationen zu verarbeiten und formallogisch exakte Schlussfolgerungen zu ziehen. Die numerische Verarbeitungskapazität ist definiert als das Heranziehen, Beziehungsstiften und Beurteilen von Informationen im Beziehungssystem Zahlen. Die verbale Verarbeitungskapazität beschreibt die Fähigkeit, im Beziehungssystem Sprache formallogisch exakt zu denken und komplexe Informationen zu verarbeiten.

Zur Erfassung der drei Inhaltsbereiche wird je ein Aufgabenbereich herangezogen. Die Auswahl des Aufgabenbereichs erfolgte daran orientiert, welcher Aufgabentyp möglichst typisch für den Inhaltsbereich ist und sich empirisch bereits gut bewährt hat. Die Bewertung der empirischen Bewährung erfolgte auch dadurch, dass möglichst Befunde zur theoretischen Fundierung der Itemkonstruktion und IRT-Skalierbarkeit der Items vorlagen.

Figurale Matrizen

Figurale Matrizen erfordern induktive Schlüsse, die nach Carpenter, Just & Shell (1990) in drei Schritten beschrieben werden können. Zunächst wird die dargestellte Matrize enkodiert, indem die Gruppen gemeinsamer und zugehöriger Elemente identifiziert werden. Anschließend werden

die Regelhaftigkeiten, nach denen die Elemente sich verändern, induziert. Im dritten Schritt werden durch die Anwendung der Regeln auf die Elementgruppen Antworten generiert bzw. Antworten gewählt. Figurale Matrizen zeigen in Studien mit die höchsten Ladungen auf einem generellen Faktor kognitiver Fähigkeiten und haben sich für die Messung von allgemeiner Intelligenz bewährt (Wilhelm & Engle, 2005). Ein Vorteil dieses Itemtyps ist darüber hinaus, dass das Beziehungssystem, in dem logische Schlüsse abgeleitet werden müssen, nicht gelernt werden kann, d.h. die Bearbeitung ist fast unabhängig von Sprache und Wissen.

Die Gestaltung der dgp-Matrizen basiert auf den Prinzipien von Becker, Preckel, Karbach, Raffel & Spinath (2014). Ähnliche Gestaltungsprinzipien, die teilweise ebenfalls einfließen, lassen sich in Carpenter et al. (1990) und Arendasy und Gittler (2003) finden. Die Items wurden vollautomatisch mit einem Item Generator in R erzeugt. Anschließend wurden sie inhaltlich geprüft und empirisch an einer relevanten Stichprobe unter high-stake Bedingungen geprüft.

Zahlenreihen

Auch Zahlenreihen erfordern induktive Schlüsse: Auf Basis einer gegebenen Reihe von Zahlen werden Regelhaftigkeiten in den Beziehungen zwischen den Zahlen abgeleitet. Diese Regelhaftigkeiten werden herangezogen, um ein fehlendes Element in der Zahlenreihe zu ergänzen.

Die Zahlenreihen unterscheiden sich durch die Eigenschaften mathematische Operationen, Größe der mathematischen Operation, Regelkomplexität (Anzahl der Regeln) und Periodizität (Länge der Reihe) (Holzman et al., 1983).

Die Items wurden vollautomatisch erzeugt und vor einer empirischen Prüfung zunächst inhaltlich geprüft. Die empirische Prüfung erfolgte unter high-stake Bedingungen.

Textanalyse

Bei der Textanalyse geht es um das Verständnis von Textinhalten. Es handelt sich hierbei um die in der schulischen Bildung vermittelte Kompetenz, mit Texten umzugehen, Texte zu verstehen und zu nutzen (Kultusministerkonferenz, 2014). Diese allgemeine Kompetenz kann nur in Teilen im Rahmen eines Testverfahrens erfasst werden. Der Fokus der Aufgaben hier liegt auf dem Verstehen von Sach- und Gebrauchstexten. In Bezug auf das Verstehen der Texte geht es insbesondere um (1) das Verständnis des Inhaltes eines Textes und das Wiedererkennen dieses Inhaltes in einer abgeänderten Formulierung und (2) das Verständnis des Inhaltes eines Textes und der Beurteilung möglicher begründeter Schlussfolgerungen aus diesem Inhalt.

Die Aufgaben zur Textanalyse beinhalten Sachtexte mit nicht-fiktionalem Inhalt. Die Texte sind in ihrer Funktion informierend (z.B. Artikel), appellierend (z.B. Reden), regulierend (z.B. Verträge) oder instruierend (z.B. Gebrauchsanweisung).

Für die Erstellung der Items liegt ein Handbuch vor. Eine automatische Erstellung ist nicht möglich. Alle Items wurden inhaltlich und empirisch unter high-stake Bedingungen überprüft.

Objektivität

Die Durchführungsobjektivität ist durch standardisierte Instruktionen gegeben. Die Instruktion und Durchführung findet am Tablet oder Computer statt. Die Bearbeitung der Items ist zeitlich begrenzt. Jeder neue Testbereich wird mit Hilfe von Beispielaufgaben schriftlich erklärt.

Die Auswertung findet voll automatisiert nach vorher definierten Auswertungsmaßstäben (Normen) statt. Auch die Interpretation erfolgt standardisiert mit Hilfe von festgelegten Eignungsstufen.

Normierung

Die Normierung erfolgt mit der Berechnung der Prozentwerte, die anschließend in Standardwerte mit einem Mittelwert von M = 100 und einer Standardabweichung von SD = 10 umgerechnet wurden. Die Normierung erfolgte für jeden Testbereich an Bewerber*innen für den mittleren und gehobenen Öffentlichen Dienst. Normen liegen im Umfang von N = 899 (FM), N = 887 (ZR) und N = 395 (TA) Jugendlichen und Erwachsenen im Alter von 16 bis 62 Jahren vor. 62 % (FM), 56 % (ZR) und 39 % (TA) der Stichprobe ist weiblich und 6 % (FM), 5 % (ZR) und 11% (TA) machten keine Angabe.

Die letzte Aktualisierung der Normen erfolgte im Jahr 2020.

Zuverlässigkeit

Die internen Konsistenzen der Testbereiche wurden mit Cronbachs Alpha geschätzt. Sie liegen für die Anzahl der Items in einem zufriedenstellenden Bereich mit Cronbachs α = .79 für 21 Zahlenreihen (numerische Verarbeitungskapazität), Cronbachs α = .72 für 18 Figurale Matrizen (figurale Verarbeitungskapazität) und Cronbachs α = .68 für 10 Textanalyse Aufgaben (verbale Verarbeitungskapazität). Die interne Konsistenz der Gesamttestbatterie kann erst berechnet werden, wenn Daten zum gesamten Testverfahren vorliegen. Eine vorläufige, konservative Schätzung der Reliabilität der Gesamttestbatterie kommt zu Werten von Cronbachs α = .81 bis Cronbachs α = .88.

Gültigkeit

Inhaltsvalidität

Die Struktur des Tests ist auf Basis des Berliner Intelligenz Strukturmodells theoretisch fundiert. Damit der Test trotz hoher Reliabilität kurz bleibt, wurde auf eine inhaltliche Breite verzichtet. Stattdessen wurde der Fokus auf die drei Aufgabentypen gelegt. Für die drei Inhaltsbereiche wurden drei typische Aufgaben auf kognitiven Leistungstests ausgewählt, die sich sowohl theoretisch als auch empirisch für die Erfassung dieser Fähigkeit gut bewährt haben (Jäger, 1970).

Konstruktvalidität

Die Überprüfung der Konstruktvalidität wurde empirisch für jeden Aufgabenbereich vorgenommen. Dafür wurde (1) die angenommene Eindimensionalität jedes Testbereichs mit Hilfe einer konfirmatorischen Faktorenanalyse überprüft, (2) die Passung des 1pl (Rasch) Modells mit Hilfe von Modelltests überprüft und (3) die Korrelationen zur bereits bewährten Testbatterie E3 im Sinne der konvergenten Validität berechnet.

Mit dem Ziel, die Daten im Sinne der Parsimonität möglichst sparsam zu modellieren, wurde die Passung der Daten an das Rasch (1pl) Modell überprüft. Fehlende Werte waren im vorliegenden Datensatz nicht vorhanden. Die Parameterschätzungen wurden mit dem R Paket eRm durchgeführt. Da nicht alle im Folgenden präsentierten Modelltests in dem Paket eRm implementiert sind, wurden Modelltests mit weiteren R-Paketen (irtoys, lavaan, mirt, psych, ltm) durchgeführt.

Die Eindimensionalität der Items wird mit einer konfirmatorischen Faktorenanalyse überprüft. Für die Überprüfung der Modellgeltung (Subgruppeninvarianz) wurden der

Andersen-Likelihood-Ratio-Test, der Wald Test und der graphische Modelltest verwendet. Des Weiteren wurde die Itemhomogenität mit dem Martin-Löf-Test überprüft. Für nähere Informationen zu den Modelltests siehe Koller, Alexandrowicz und Hatzinger (2012). Auch nach Koller et al. (2012) wird das globale Signifikanzniveau der Tests auf α = .10 festgelegt. Eine Alphakorrektur auf Grund von multiplen Tests wird innerhalb der Gruppen von Modelltests durchgeführt (siehe S. 168, Koller et al., 2012). Daher unterscheidet sich das Signifikanzniveau α bei den verschiedenen Modelltests.

Figurale Matrizen

Dimensionalität

Zur Überprüfung der Eindimensionalität wurde eine konfirmatorische Faktorenanalyse auf Itemebene durchgeführt (MLM, X^2 =211.16, df = 135, p = .000, CFI = .931, RMSEA = .025, 90% CI [.018; .031]). Der Modellfit kann auf Grundlage der Fit Indices als zufriedenstellend bewertet werden.

Subgruppeninvarianz

Zur Überprüfung der Itemhomogenität wurde als globaler Fit-Index der Andersen-Likelihood-Ratio Test (Andersen, 1973) verwendet. Hier werden die Likelihoods getrennt für die Gesamtgruppe und zwei Teilgruppen (geteilt am Median) berechnet und miteinander verglichen. Bei Modellgeltung (H0) sollten sich diese Likelihoods nicht voneinander unterscheiden. Wird der LRT Test signifikant, ist die Modellgeltung nicht gegeben. Für die selektierten Items wird der Anderson-LRT Test nicht signifikant (LR-value = 25.471, X^2 df = 17, p-value = .085) und die Annahme der Modellgeltung kann beibehalten werden. Hier wird das Signifikanzniveau lokal auf α = .05 festgelegt.

Auf Itemebene kann der Wald-Test herangezogen werden, um die Modellgeltung zu überprüfen. Der Test basiert ebenfalls auf einem Likelihood-Ratio-Test. Hierbei werden die Schätzungen der Schwierigkeitsparameter in zwei Gruppen (Mediansplit) miteinander verglichen. Diese sollten nicht voneinander abweichen. Unter Berücksichtigung der Alpha Korrektur für multiples Testen wird das lokale Signifikanzniveau auf α = .003 festgelegt. Keines der Items zeigt signifikante Unterschiede in den Schwierigkeitsparameterschätzungen für die am Median getrennten Gruppen.

Ebenfalls kann auf Itemebene ein statistischer Ansatz zur Überprüfung des Itemfits herangezogen werden (Reise, 1990). Hier werden modell-implizierte und empirische IRCs miteinander verglichen. Signifikante Ergebnisse (α = .003) zeigen problematische Items an. Auch hier können bei keinem Item signifikante Abweichungen beobachtet werden.

Die Grafische Modellkontrolle (siehe Abbildung 1) zeigt ein Streudiagramm der Parameterschätzungen für die Items aus zwei Personengruppen (Mediansplit). Das lokale Signifikanzniveau wurde auch hier auf α = .003 für die Konfidenzintervalle (rote Ellipsen) festgelegt. Keines der noch vorhandenen Items zeigt signifikante Abweichungen.

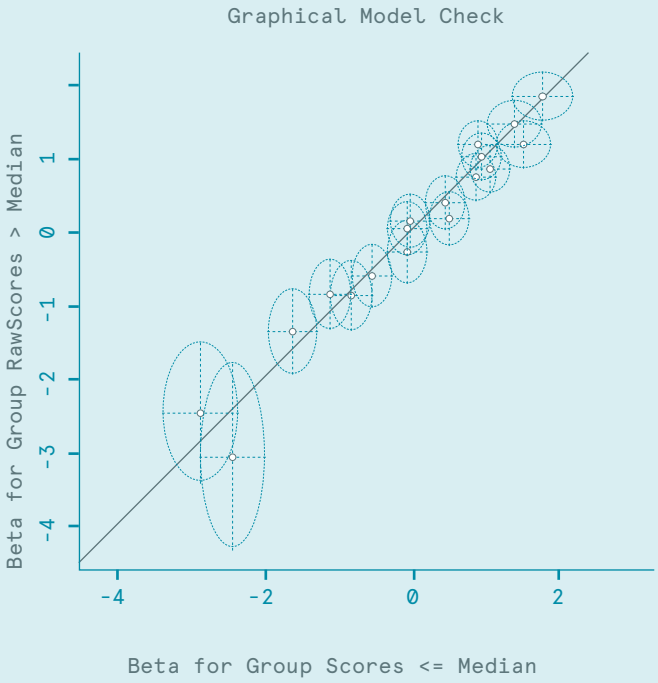

Abbildung 1. Graphischer Modelltest für die Figuralen Matrizen

Itemhomogenität

Beim Martin-Löf-Test werden die Likelihoods für zwei Itemstichproben miteinander verglichen. Bei Geltung des Rasch-Modells sollten sich die Likelihoods für zwei Itemgruppen (Mediansplit) nicht unterscheiden, da diese dasselbe messen sollten. Der Martin-Löf-Test wird für die vorliegenden Daten nicht signifikant (LR-value = 59.327, X2 df = 80, p-value = .96) bei einem lokalen Signifikanzniveau von α = 0.05. Itemhomogenität ist demnach gegeben.

Korrelationen zum E3

Auf latenter Ebene zeigen sich moderate bis hohe Korrelationen zwischen den Figuralen Matrizen und den E3 Subtests (siehe Tabelle 1), wie sie auf Grundlage der Intelligenztheorie sowie empirischen Befunden zu erwarten sind. Der Fit für dieses Modell ist sehr gut (ML, X2 =228.42, df = 142, p = .000, CFI = .986, RMSEA = .026, 90% CI [.020; .032]).

	AG	KL	ZR	TX
FM	.508	.466	.525	.587

Tabelle 1. Korrelationen auf latenter Ebene

Zahlenreihen

Dimensionalität

Zur Überprüfung der Dimensionalität wurde eine konfirmatorische Faktorenanalyse durchgeführt (MLM, X2 =381.485, df = 189, p = .000, CFI = .904, RMSEA = .034, 90% CI [.029;.039]). Der Modellfit kann als zufriedenstellend bewertet werden.

Subgruppeninvarianz

Für die selektierten Items wird der Anderson-LRT-Test (Andersen, 1973) nicht signifikant (LR-value = 29.074, X2 df = 20, p-value = .086) und die Annahme der Modellgeltung kann beibehalten werden. Hier wird das lokale Signifikanzniveau auf α = .05 festgelegt.

Auf Itemebene kann der Wald-Test herangezogen werden, um die Modellgeltung zu überprüfen. Unter Berücksichtigung der Alpha Korrektur für multiples Testen wird das lokale Signifikanzniveau auf α = .002 festgelegt. Keines der Items zeigt signifikante Unterschiede in den Schwierigkeitsparameterschätzungen für die am Median getrennten Gruppen.

Ebenfalls kann auf Itemebene ein statistischer Ansatz (BCHI) zur Überprüfung des Itemfits herangezogen werden (Reise, 1990). Signifikante Ergebnisse (α = .002) zeigen problematische Items an. Auch hier konnten bei keinem Item signifikante Abweichungen beobachtet werden.

Die grafische Modellkontrolle (siehe Abbildung 2) zeigt ein Streudiagramm der Parameterschätzungen für die Items aus zwei Personengruppen (Mediansplit). Das lokale Signifikanzniveau wurde auch hier auf α = .002 für die Konfidenzintervalle (rote Ellipsen) festgelegt. Keines der noch vorhandenen Items zeigt signifikante Abweichungen.

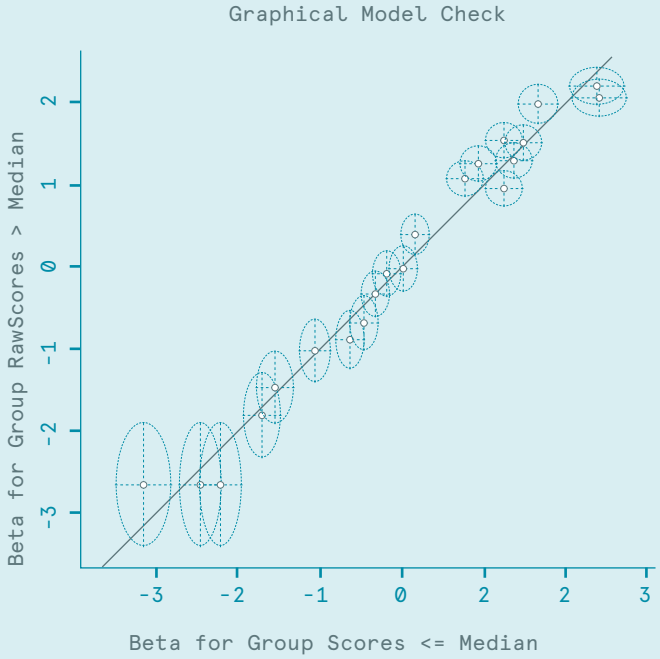

Abbildung 2. Graphischer Modelltest für die Zahlenreihen

Itemhomogenität

Der Martin-Löf-Test wird für die vorliegenden Daten nicht signifikant (LR-value = 71.427, X2 df = 109, p-value = .998) bei einem lokalen Signifikanzniveau von α = .05. Itemhomogenität ist demnach gegeben.

Korrelationen zum E3

Auf latenter Ebene zeigen sich moderate bis hohe Korrelationen zwischen den latenten Variablen der Subtests (siehe Tabelle 2). Wie erwartet, korrelieren Zahlenreihen aus dem E3 und die neuen Zahlenreihen sehr hoch miteinander. Der Fit für dieses Modell ist sehr gut (ML, X2 =221.629, df = 142, p = .000, CFI = .987, RMSEA = .025, 90% CI [.019;.031]).

	AG	KL	ZR	TX
ZR	.558	.556	.828	.722

Tabelle 2. Korrelationen auf latenter Ebene

Textanalyse

Dimensionalität

Zur Überprüfung der Dimensionalität wurde eine konfirmatorische Faktorenanalyse durchgeführt (MLM, X2 =44.682, df = 35, p = .126, CFI = .996, RMSEA = .026, 90% CI [.000;.046]). Der Modellfit kann als sehr gut bewertet werden.

Subgruppeninvarianz

Für die selektierten Items wird der Anderson-LRT Test (Andersen, 1973) nicht signifikant (LR-value = 5.082, X2 df = 7, p-value = .65) und die Annahme der Modellgeltung kann beibehalten werden. Hier wird das lokale Signifikanzniveau auf α = .05 festgelegt.

Auf Itemebene kann der Wald-Test herangezogen werden, um die Modellgeltung zu überprüfen. Unter Berücksichtigung der Alpha Korrektur für multiples Testen wird das lokale Signifikanzniveau auf α = .005 festgelegt. Keines der Items zeigt signifikante Unterschiede in den Schwierigkeitsparameterschätzungen für die am Median getrennten Gruppen.

Ebenfalls kann auf Itemebene ein statistischer Ansatz (BCHI) zur Überprüfung des Itemfits herangezogen werden (Reise, 1990). Signifikante Ergebnisse (α = .005) zeigen problematische Items an. Auch hier konnten bei keinem Item signifikante Abweichungen beobachtet werden.

Die grafische Modellkontrolle (siehe Abbildung 3) zeigt ein Streudiagramm der Parameterschätzungen für die Items aus zwei Personengruppen (Mediansplit). Das Signifikanzniveau wurde auch hier auf α = .005 für die Konfidenzintervalle (rote Ellipsen) festgelegt. Keines der noch vorhandenen Items zeigt signifikante Abweichungen.

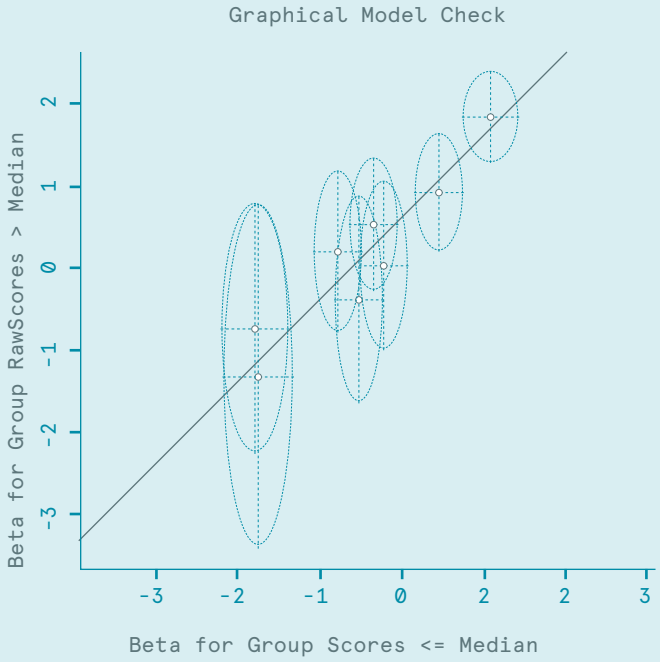

Abbildung 3. Graphischer Modelltests für die Textanalyse Items

Itemhomogenität

Der Martin-Löf-Test wird für die vorliegenden Daten nicht signifikant (LR-value = 21.897, X2 df = 24, p-value = .585). Itemhomogenität ist demnach gegeben.

Korrelationen zum E3

Auf latenter Ebene zeigen sich moderate bis hohe Korrelationen zwischen den latenten Variablen der Subtests (siehe Tabelle 3). Die Korrelationen der Textanalyse Items zu den verbalen Bereichen des E3 sind dabei wie erwartet höher als zu den numerischen Bereichen. Der Fit für dieses Modell ist sehr gut (MLM, X2 =314.660, df = 265, p = .020, CFI = .983, RMSEA = .022, 90% CI [.010;.031]).

	AG	KL	ZR	TX
TA	.734	.728	.462	.599

Tabelle 3. Korrelationen auf latenter Ebene

Einordnung der Ergebnisse

Die Evidenz für die Konstruktvalidität aller Subtests kann als sehr gut bewertet werden. Die Items jedes Subtests sind eindimensional. Alle Modelltests sprechen für die Modellgeltung des Raschmodells für die jeweiligen Subtests. Die konvergente Validität wird durch die Korrelationen zu den Subtests des E3 gestützt.

Prädiktive Validität

Auf Grundlage der zahlreichen Evidenz, die es bezüglich der Vorhersagekraft kognitiver Fähigkeiten für Ausbildungs- und Berufserfolg gibt (für einen Überblick siehe Ones, Viswesvaran & Dilchert, 2005), besteht die Möglichkeit der Validitätsgeneralisierung (Kersting, 2009). Das bedeutet, dass bei bestätigter Konstruktvalidität bisherige Befunde zur prädiktiven Validität auf das in Frage stehende Verfahren bezogen werden können. Die vorliegenden Aufgaben sind sowohl inhaltlich als auch theoretisch fundierte Aufgaben zur Messung von kognitiven Fähigkeiten, die empirisch im Rahmen

der Konstruktvalidität (siehe oben) ausführlich überprüft wurden. Man kann also auch für den VK-1 eine prädiktive Validität für Ausbildungs- und Berufserfolg annehmen, die sich im Rahmen dessen bewegt, was in Meta-Analysen an Vorhersageerfolg gezeigt werden kann. Sobald Daten für die Vorhersage von Erfolgskriterien vorliegen, wird diese Annahme überprüft.

Weitere Gütekriterien

Fairness

Figurale Matrizen

Im Durchschnitt lösen Männer (N = 293) M = 10.83 (SD = 3.547) und Frauen (N = 555) M = 10.33 (SD = 3.233) FM Aufgaben korrekt.

Zur Beurteilung von uniformen und non-uniformen Differential Item Functioning wurden verschiedene Statistiken herangezogen. Die Mantel-Haenszel-Statistik (uniform, non IRT) (Holland & Thayer, 1988) zeigt für kein Item einen signifikanten Wert. Die entsprechenden Effektgrößen fallen weitestgehend klein aus und sind für ein Item moderat bzw. stark. Laut der Logistischen Regression (non-uniform, non-IRT) (Swaminathan and Rogers, 1990) zeigen zwei Items leichte Auffälligkeiten, wobei alle Effektgrößen vernachlässigbar sind.

Für den Vergleich der Geschlechtsgruppen zeigt sich beim Likelihood-Ratio-Test (uniform, IRT) (Thissen, Steinberg & Wainer, 1988) hier auch ein nicht signifikantes Ergebnis mit (LR-value = 33.383, X2 df = 34, p-value = .498). Die Lords Chi-Quadrat Statistik (non-uniform, IRT) (Lord, 1980) wird für kein Item signifikant. Weitestgehend liegen kleine Effektgrößen vor, mit Ausnahme von zwei moderaten Effekten.

Zahlenreihen

Im Durchschnitt lösen Männer (N = 345) M = 9.525 (SD = 3.102) und Frauen (N = 499) M = 8.491 (SD = 3.154) ZR Aufgaben korrekt.

Zur Beurteilung von uniformem und non-uniformem Differential wurden verschiedene Statistiken herangezogen. Die Mantel-Haenszel Statistik (uniform, non IRT) (Holland & Thayer, 1988) zeigt für kein Item einen signifikanten Wert. Die entsprechenden Effektgrößen fallen alle klein aus. Auch laut der Logistischen Regression (non-uniform, non-IRT) (Swaminathan and Rogers, 1990) zeigt kein Item Auffälligkeiten und alle Effektgrößen sind vernachlässigbar.

Für den Vergleich der Geschlechtsgruppen zeigt sich beim Likelihood-Ratio-Test (uniform, IRT) (Thissen, Steinberg & Wainer, 1988) hier auch ein nicht signifikantes Ergebnis (LR-value = 11.551, X2 df = 20, p-value = .931). Die Lords Chi-Quadrat Statistik (non-uniform, IRT) (Lord, 1980) wird ebenfalls für kein Item signifikant und es liegen durchgehend kleine Effekte vor.

Textanalyse

Im Durchschnitt lösen Männer (N = 196) M = 7.17 (SD = 2.12) und Frauen (N = 155) M = 7.42 (SD = 2.08) TA Aufgaben korrekt.

Zur Beurteilung von uniformem und non-uniformem Differential wurden verschiedene Statistiken herangezogen. Die Mantel-Haenszel-Statistik (uniform, non IRT) (Holland & Thayer, 1988) zeigt für kein Item einen signifikanten Wert. Die entsprechenden Effektgrößen fallen nur für zwei Items moderat aus. Auch laut der Logistischen Regression (non-uniform, non-IRT) (Swaminathan and Rogers, 1990) zeigt kein Item Auffälligkeiten und alle Effektgrößen sind vernachlässigbar.

Für den Vergleich der Geschlechtsgruppen zeigt sich beim Likelihood-Ratio-Test (uniform, IRT) (Thissen, Steinberg & Wainer, 1988) hier auch ein nicht signifikantes Ergebnis (LR-value = 11.551, X2 df = 20, p-value = .931). Die Lords Chi-Quadrat Statistik (non-uniform, IRT) (Lord, 1980) wird ebenfalls für kein Item signifikant und es liegen durchgehend kleine Effekte vor.

Ökonomie

Der VK-1 ist ein sehr kurzer Test, der dennoch einen hohen Erkenntnisgewinn dadurch erbringt, dass die testtheoretische Fundierung des Verfahrens hohen diagnostischen Ansprüchen gerecht wird.

Skalierung

Nach Moosbrugger und Kelava (2008) erfüllt ein Test dann das Gütekriterium der Skalierung, wenn die „laut Verrechnungsregel resultierenden Testwerte die empirischen Merkmalsrelationen adäquat abbilden“ (S. 18). Dieses Kriterium gilt unter anderem dann als gegeben, wenn für ein Testverfahren die Geltung des Raschmodells gezeigt werden konnte (Schmidt-Atzert & Amelang, 2012). Für alle Subtests des VK-1 konnte Evidenz für die Geltung des Rasch-Modells gesammelt werden.

Fazit

Die kurze Testbatterie VK-1 ermöglicht die sehr ökonomische und faire Erfassung grundlegender kognitiver Fähigkeiten im Rahmen einer multimodalen Auswahlstrategie, beispielsweise als eine Station im Assessment-Center. Die Durchführung, Auswertung und Interpretation ist durch standardisierte Bedingungen gesichert. Die Normen der Testbereiche sind aktuell und wurden auf Basis von Bewerber*innen des Öffentlichen Dienstes unter high-stake Bedingungen (reale Bewerbungssituation) erhoben. Die Reliabilität der Testbereiche ist – insbesondere unter Berücksichtigung der Testlänge – als gut zu bewerten. Die Inhaltsvalidität ist durch die theorie-geleitete, teils automatisierte Konstruktion der Aufgaben als gut zu bewerten. Es liegt eine vielversprechende Evidenz zur Konstruktvalidität vor, insbesondere in Bezug auf die faktorielle Validität (Eindimensionalität), die Geltung des Rasch-Modells sowie die konvergente Validität. Annahmen zur prädiktiven Validität des Testverfahrens können auf Grundlage der Validitätsgeneralisierung vorgenommen werden.

Die Qualität des VK-1 wird in zukünftigen Untersuchungen fortlaufend überprüft: Die Normen werden nach ungefähr einem Jahr geprüft, eine Erweiterung der Normen, getrennt für den mittleren und gehobenen Dienst, wird angestrebt und eine Bewährungskontrolle zur Untersuchung der prädiktiven Validität wird baldmöglichst durchgeführt.

Literatur:

Andersen, E.B. (1973). A goodness of fit test for the Rasch model. Psychometrika, 38, 123-140. DOI: 10.1007/bf02291180

Arendasy, M. E. und Gittler, G. (2003). IRT-basierter Vergleich zweier Varianten automatisiert erstellter Matrizentestaufgaben. Zeitschrift für Differentielle und Diagnostische Psychologie, 24, 261-175. DOI 10.1024//0170-1789.24.4.261

Becker, N., Preckel, F., Karbach, J., Raffel, N. & Spinath, F. M. (2014). Die Matrizenkonstruktionsaufgabe: Validierung eines distraktorfreien Aufgabenformats zur Vorgabe figuraler Matrizen. Diagnostica, 61, 22-33. DOI: 10.1026/0012-1924/a000111

Carpenter, P. A., Just, M. A. & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test. Psychological Review, 97, 404-431. DOI: 10.21236/ada221448

Holzman, T. G., Pellegrino, J. W. & Glaser, R. (1983). Cognitive variables in series completion. Journal of Educational Psychology, 75, 603-618. DOI: 10.1037/0022-0663.75.4.603

Jäger, A. O., Süß, H.-M., & Beauducel, A. (1997). BIS-TEST Berliner Intelligenzstruktur-Test – Form 4. Göttingen: Hogrefe.

Jäger, A. O. (1970). Dimensionen der Intelligenz. Göttingen: Hogrefe.

Kersting, M. (2009). Die DGP Testverfahren - Ein kurzer Rückblick und eine aktuelle Studie zur Konstrukt und Kriteriumsvalidität des BIS-r-DGP Tests. dgp Informationen, 51, 22-37.

Kultusministerkonferenz, 2014. Integriertes Kompetenzstufenmodell zu den Bildungsstandards für den Hauptschulabschluss und den Mittleren Schulabschluss im Fach Deutsch für den Kompetenzbereich. Beschluss der Kultusministerkonferenz (KMK) vom 11.12.2014.

Koller, I., Alexandrowicz, R. & Hatzinger, R. (2012). Das Rasch-Modell in der Praxis: Eine Einführung mit eRm. Wien: Facultas.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates. DOI: 10.4324/9780203056615

Moosbrugger, H. und Kelava, A. (2008) Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In Helfried Moosbrugger & Augustin Kelava (Hrsg.), Testtheorie und Fragebogenkonstruktion (pp. 7-25). Heidelberg: Springer. DOI: 10.1007/978-3-540-71635-8

Ones, D. S., Viswesvaran, C. & Dilchert, S. (2005). Cognitive Ability in Selection Decisions. In Oliver Wilhelm & Randall W. Engle, Handbook of Understanding and Measuring Intelligence (pp. 431-468). Thousand Oaks: Sage Publications. DOI: 10.4135/9781452233529

Reise, S. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. Applied Psychological Measurement, 14, 127–137. DOI: 10.1177/014662169001400202

Schmidt-Atzert, L. & Amelang, M. (2012). Psychologische Diagnostik. Heidelberg: Springer. DOI: 10.1007/978-3-642-17001-0

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361–370. DOI: 10.1111/j.1745-3984.1990.tb00754.x

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), Test validity (p. 147–172). Lawrence Erlbaum Associates, Inc.

Wilhelm, O. & Engle, R. W. (2005). Introduction: Intelligence: A diva and a workhorse. In Oliver Wilhelm & Randall W. Engle, Handbook of Understanding and Measuring Intelligence (pp. 1-10). Thousand Oaks: Sage Publications. DOI: 10.4135/9781452233529

Kontakt:

Dipl.-Psych. Dr. Anna-Lena Jobmann
jobmann@dgp.de

Deutsche Gesellschaft für Personalwesen e. V.
Kantstr. 153, 10623 Berlin