

Petri, P.S., Weingardt, V. & Kersting, M. (2019). Let's get to the hard facts – Erfassung von Intelligenz im Rahmen von Online Self-Assessments. *Empirische Pädagogik*, 33, 307–330.

Empirische Pädagogik

33. Jahrgang 2019

3. Heft

Empirische Pädagogik
2019, 33. Jahrgang, Heft 3, S. 307-330

© Verlag Empirische Pädagogik

Pascale Stephanie Petri, Vera Weingardt & Martin Kersting

Let's get to the hard facts – Erfassung von Intelligenz im Rahmen von Online Self-Assessments

Zahlreiche, auch meta-analytische Befunde belegen, dass Intelligenz der beste Prädiktor für Leistung sowohl im beruflichen (Schmidt & Hunter, 1998, 2004; Schmidt, Shaffer & Oh, 2008) als auch im Studienkontext (Richardson, Abraham & Bond, 2012) ist. Im Rahmen von Online Self-Assessments (OSA), die viele Hochschulen Studieninteressierten als Selbsttest- und Informationstools für eine fundierte Studienwahl zur Verfügung stellen, kommen Intelligenztests dennoch eher selten zum Einsatz. Das hat hauptsächlich drei Gründe: (1) Die Durchführung dieser Tests erfordert oft einen erheblichen zeitlichen Aufwand. (2) Zudem ist die Bearbeitung eines Leistungstests im Vergleich zu den häufig verwendeten Interessens- und Persönlichkeitsfragebogen belastender: Die Bearbeitung kostet mehr kognitive Anstrengung. Zusätzlich kann Frust entstehen, wenn Nutzer bemerken, dass sie Aufgaben nicht lösen können. Dies kann in freiwilligen OSA Nutzungsabbrüche nach sich ziehen. (3) Ein dritter Grund ist die Verfügbarkeit. Während im Bereich der Interessens- und Persönlichkeitsfragebogen relativ viele Verfahren mit freien Lizenzen zur Verfügung stehen, ist dies im Bereich Fähigkeitstests seltener der Fall. Dies führt dazu, dass für die Nutzung in OSA aufwändige finanzielle Investitionen oder Testentwicklungen notwendig sind. Will man also den besten Prädiktor für Studienleistungen in OSA nutzen, müssen diese Herausforderungen überwunden werden. Wir haben einen Fähigkeitstest entwickelt, der sich in der Online-Durchführung bereits bewährt hat und durch eine kurze Bearbeitungszeit (unter 30 Min.) überzeugt. Erfasst werden verbale, numerische und figurale Verarbeitungskapazität gemäß dem Berliner Intelligenzstrukturmodell (Jäger, 1984). Im vorliegenden Artikel berichten wir Befunde zur Prüfung der Reliabilität, Validität sowie zu den Itemgütekriterien. Der Test steht auf Anfrage für die Nutzung in OSA kostenfrei zur Verfügung.

Schlagwörter: Kognitive Fähigkeiten – Online Self-Assessment – Reliabilität – Validität

Let's get to the hard facts – Assessing cognitive ability in online self-assessments

There is strong, even meta-analytic evidence, that cognitive abilities are the best predictor for performance not only in vocational (Schmidt & Hunter, 1998, 2004; Schmidt et al., 2008) but also in educational settings (Richardson et al., 2012). Many universities offer so-called online self-assessments (OSA) to future students as a tool for self-assessment and as an information platform to help them arriving at an informed decision whether and what to study. However, in this counseling context the assessment of cognitive abilities is rare. This is mainly due to three reasons: (1) Cognitive ability tests are often time-costly. (2) Compared to the widely used personality and interest questionnaires, taking a cognitive ability test is more straining for the users because it requires high cognitive effort and could be frustrating for underperformers. This in turn could provoke dropout. In addition, (3) there are less cognitive ability tests than personality or interest questionnaires that come with an open content. Thus, implementing the best predictor of student performance in OSA requires high financial investments or a lot of test construction work. To overcome these challenges, we developed a cognitive ability test that has already proven in an online setting and can be administered in less than 30 min. Our test assesses verbal, numeric and spatial processing capacity in the sense of the

Berlin Intelligence Structure Model (Jäger, 1984). This paper addresses our findings regarding reliability and validity as well as item statistics. On demand, we provide the test for implementation in OSA.

Keywords: cognitive ability – online self-assessment – reliability – validity

1 Einleitung

1.1 Erfolg vs. Abbruch im Studium

In Deutschland brechen im Mittel über alle Studienfächer und Hochschularten hinweg circa 29 % aller Studienanfänger ihr Studium ab (Heublein et al., 2017). Studienabbruch verursacht Kosten auf verschiedenen Ebenen (Isleib, 2015). Auf der Makroebene, also aus gesamtgesellschaftlicher Perspektive, ist es so, dass Studienabbrecher zwar eine Zeit lang im Hochschulsystem verweilen, jedoch nicht die abschließende Qualifikation erwerben, durch welche sich die staatliche Förderung vieler Hochschulen letztendlich für die Gesellschaft auszahlen könnte. Hinzu kommt auf der Mesebene, dass für viele Hochschulen ein Teil der Mittelvergabe auch an die Anzahl erfolgreicher Studienabsolventen gebunden ist. Jeder Studienabbrecher bedeutet für die Hochschulen folglich einen finanziellen Verlust. Auf der Mikroebene steht das Individuum im Fokus: Bricht ein Studierender sein Studium ab, so ist davon auszugehen, dass er unter anderem Kosten auf finanzieller und emotionaler Ebene hat. Es lässt sich zusammenfassend festhalten, dass eine Reduktion der Studienabbruchquote erstrebenswert ist. Eine prophylaktische Intervention auf institutioneller Ebene ist es, Studieninteressierte ausführlich über Studienanforderungen im Allgemeinen und die individuelle Studieneignung im Speziellen zu informieren (Wissenschaftsrat, 2004). Dies soll langfristig die Studienabbruchquote reduzieren, indem ungünstige Studien(fach)wahlen weniger wahrscheinlich werden.

1.2 Online Self-Assessments im Hochschulkontext

Eine Möglichkeit, Studieninteressierte mit eben diesen Informationen zu versorgen, bieten Online Self-Assessment (OSA). Das sind Online-Tools, die Fragebogen, Tests und Informationsmodule zur selbstständigen Nutzung zur Verfügung stellen. Im Kontext tertiärer Bildung werden OSA oft an Hochschulen entwickelt und betrieben, um Studieninteressierten die Möglichkeit zu geben, ihre individuelle Studieneignung und/oder Passung zum etwaigen Wunschstudienfach zu prüfen (Hasenberg & Schmidt-Atzert, 2014; Heukamp, Putz, Milbradt & Hornke, 2009). Darüber hinaus können sie sich über das Studium im Allgemeinen, die jeweilige Hochschule und/oder ein bestimmtes Fach an einer bestimmten Hochschule im Speziellen informieren. OSA-Angebote sollen eine fundierte Studien(fach)wahl ermöglichen. Dazu wird nach Bearbeitung des OSA ein ausführliches automatisches Feedback erstellt, um den Studieninteressierten zurückzumelden, inwiefern beispielsweise ihre studienbezogenen Interessen, Eigenschaften, Fähigkeiten und ihr Vorwissen zu den

Charakteristiken des Studiengangs oder zum Hochschulstudium allgemein passen (Heukamp et al., 2009). Die tiefgreifende Auseinandersetzung mit den allgemeinen Anforderungen und mit Studieninhalten bestimmter Fächer soll helfen, die eigene Eignung und den bereits vorhandenen Wissensstand zu reflektieren. Etwaige Fehlerwartungen können durch ausführliche Informationen korrigiert werden. Dies soll letztendlich die Anzahl an Studienabbrüchen reduzieren im Sinne vorbeugender Studien(fach)wahlberatung. Daneben sind auch die Personen zu bedenken, die sich ein Studium – z. B. aufgrund eher unterdurchschnittlicher Schulleistungen oder aufgrund ihrer sozialen Herkunft – trotz vorhandener Eignung nicht zutrauen und daher auf ein Studium verzichten. Solche „underachiever“ (Whitmore, 1980) haben im Rahmen von OSA die Möglichkeit, ohne Leistungsdruck ihre Fähigkeiten und Eigenschaften mit den Anforderungen abzugleichen und werden durch das Feedback ermutigt, ein Studium aufzunehmen.

1.3 Ausrichtung von OSA

Das Spektrum an Inhalten von OSA ist relativ breit. Grundsätzlich kann man unterscheiden, auf welchem Generalitäts- vs. Spezifitätslevel ein konkretes OSA angesiedelt ist (Heukamp et al., 2009). OSA zur Überprüfung der allgemeinen Studierfähigkeit dienen zur grundsätzlichen Klärung, ob man sich für ein Hochschulstudium eignet. Fachspezifische OSA fokussieren – meist zusätzlich – konkrete Studieninhalte in einzelnen Fächern oder Fachgruppen. Diese OSA sind am häufigsten zu finden (Hochschule der Bundesagentur für Arbeit und der Fachhochschule Nordwestschweiz, 2018).

1.4 Inhalte von OSA

Vorangegangene Forschung hat diverse valide Prädiktoren von sogenannten harten Kriterien des Studienerfolgs (i. d. R. Noten) aufgezeigt. Darunter Interesse am Fach, Gewissenshaftigkeit, Selbstwirksamkeit im Studium und insbesondere auch Intelligenz (Richardson et al., 2012). Speziell Intelligenztests haben sich als hochgradig prädiktiv für Erfolg in Studium und Beruf erwiesen (Salgado et al., 2003). Konkret für den Hochschulkontext konnte dies sogar bereits meta-analytisch nachgewiesen werden: Die Intelligenz einer Person ist einer der zuverlässigsten Prädiktoren für Studienleistung (Richardson et al., 2012). Anders herum betrachtet sind Leistungsprobleme der am häufigsten genannte Grund für einen Studienabbruch (Heublein et al., 2017). Diese entstehen u. a. durch kognitive Überforderung. Daher bietet es sich an, im Rahmen von OSA auch Intelligenzdiagnostik durchzuführen. Grundsätzlich gilt es dabei allerdings zu bedenken, dass sowohl Intelligenztests, als auch alle anderen Verfahren im Rahmen von OSA lediglich als Screening-Instrumente verstanden werden sollten. Die Kürze der Bearbeitung und das Setting des unbeaufsichtigten Testens erlauben lediglich eine erste Einschätzung. Für detaillierte

Individualdiagnostik sollten ausführliche und beaufsichtigte Testungen angeschlossen werden. Daher sind OSA als ein Element der Studienwahlberatung zu sehen, nicht als Ersatz für beispielsweise persönliche Gespräche mit Studienberatern.

1.5 Erfassung von Intelligenz in OSA

Im Rahmen von Online Self-Assessments (OSA) kommen Tests für kognitive Fähigkeiten eher selten zum Einsatz. Das hat hauptsächlich drei Gründe:

(1) Insbesondere im Rahmen freiwilliger OSA sind kurzweilige Tests und Fragebogen besonders gefragt, um die Nutzer während der Testdurchführung nicht zu verlieren. Intelligenzdiagnostik hingegen erfordert in der Regel eine relativ zeitaufwändige Testung.

(2) Etablierte deutschsprachige Tests wie der Intelligenz-Struktur-Test 2000 R (Liepmann, Beauducel, Brocke & Amthauer, 2007) oder der Wilde Intelligenztest 2 (Kersting, Althoff & Jäger, 2008) unterliegen dem Testschutz. Daher können sie nicht für OSA im Rahmen ungeschützter Testungen, online und ohne Aufsicht verwendet werden. Es gibt aber nur ein sehr kleines Angebot an wissenschaftlich fundierten Intelligenztests mit deutschsprachigen Inhalten, für die freie Lizenzen vorliegen und die für den ungeschützten Einsatz vorgesehen sind. Beispiele sind der Hagener Matrizen Test (HMT) (Heydasch, Renner, Haubrich, Hilbig & Zettler, 2013), und der BEFKI GC-K (Schipolowski et al., 2013). Beide Tests sind sehr spezifisch auf einzelne Intelligenzdimensionen ausgerichtet. Der HMT deckt aus dem Berliner Intelligenzstrukturmodell (siehe 2.1, BIS, Jäger, 1984) nur den Bereich der figuralen Verarbeitungskapazität ab, der BEFKI GC-K beschränkt sich auf die kristalline Intelligenz (Cattell, 1943). Da in den meisten Studiengängen mehrere intellektuelle Dimensionen gefragt sind, ist aufgrund des Symmetrie-Prinzip (Brunswick, 1956) davon auszugehen, dass zu spezifische Tests zur Prädiktion von Studienerfolg weniger geeignet sind.

(3) Was den Einsatz von Intelligenztests in OSA zudem erschwert, ist dass die Bearbeitung eines Leistungstests im Vergleich zu den häufig verwendeten Interessens- und Persönlichkeitsfragebogen belastender für die Nutzer sein kann. Die Bearbeitung kostet mehr kognitive Anstrengung. Zusätzlich kann Frust entstehen, wenn Nutzer bemerken, dass sie Aufgaben nicht lösen können. Dies kann in freiwilligen OSA zu Nutzungsabbrüchen führen. Trotz dieser Herausforderungen sollten – bereits aufgrund des Prinzips der Multimethodalität – für OSA nicht nur Selbstbeschreibungsinstrumente (Fragebogen, Interessenstests) genutzt werden. Besonders problematisch ist es, OSA auf Interessenstests zu beschränken. Hier besteht die Gefahr, dass lediglich Stereotype abgefragt und bedient werden, was insbesondere dazu führen kann, dass Frauen „Frauenfächer“ (etwas mit „Menschen“) und Männern „Männerfächer“ (etwas mit „Maschinen“) empfohlen werden (Franzke, 2010). Ein

weiteres Problem von Selbstberichtsfragebogen (z. B. fachspezifisch-studienbezogene Persönlichkeitsfragebogen) ist, dass diese ein gewisses Maß an Selbsterkenntnis voraussetzen. Ausschlaggebend für den Einsatz von Intelligenztests bleibt aber die Tatsache, dass diese eine deutlich höhere Kriteriumsvalidität aufweisen als Persönlichkeits- und Interessensfragebogen.

1.6 Intendierter Beitrag

Auf Grund dieser Situation haben wir einen Intelligenztest entwickelt, der die oben genannten Herausforderungen überwindet: der Gießener kognitive Kompetenzen Test, kurz GkKT (Ulfert, Ott, Bothe & Kersting, 2017). Mit einer zeitökonomischen Administration von weniger als 30 Minuten, kann die Verarbeitungskapazität (schlussfolgerndes Denken) in Bezug auf verbale, numerische und figurale Inhalte gemäß BIS-Modell (Jäger, 1984) erfasst werden. Der vorliegende Beitrag soll den Konstruktionshintergrund, den Aufbau sowie die bisher erhobenen psychometrischen Gütekriterien des Tests darstellen.

2 Theorie

2.1 Einordnung des GkKT in etablierte Intelligenztheorien

Gemäß dem BIS-Modell (Jäger, 1984) können zunächst zwei Arten von Fähigkeiten unterschieden werden: die operativen und die inhaltsgebundenen. Die vier postulierten operativen Fähigkeiten sind Verarbeitungskapazität, Einfallsreichtum, Merkfähigkeit und Bearbeitungsgeschwindigkeit. Die drei postulierten inhaltsgebundenen Fähigkeiten verbal, numerisch und figural, sind charakterisiert anhand der Art des Aufgabenmaterials. Gemäß Jäger (1984) bildet die allgemeine Intelligenz in diesem Modell das „Integral“ aller sieben genannten Fähigkeiten. Im Berliner Intelligenzstruktur-Test (Jäger, Süß & Beauducel, 1997) können die vier operativen Fähigkeiten mit den drei Inhaltsbereichen gekreuzt werden, sodass zwölf Facetten entstehen. Im hier vorgestellten GkKT wird die operative Fähigkeit der Verarbeitungskapazität erfasst, hinsichtlich aller drei Inhaltsbereiche. Dabei gibt es pro Inhaltsbereich mehrere Aufgabengruppen zur Erfassung der jeweiligen Fähigkeit. Im Sinne der Three-Stratum-Theory (Carroll, 1993) erfasst der GkKT fluide Intelligenz. Auch bei der Unterscheidung zwischen fluider und kristalliner Intelligenz nach Cattell (1943), lässt sich die mit dem GkKT erfasste Fähigkeit in die Kategorie fluide Intelligenz einordnen. Ordnet man die mit dem GkKT erfasste Fähigkeit in das Cattell-Horn-Carroll-Modell der Intelligenz (McGrew, 2009) ein, lässt sich festhalten, dass schlussfolgerndes Denken erfasst wird. Obgleich der GkKT auf Basis des BIS-Modells (Jäger, 1984) entwickelt wurde, möchten wir im Folgenden statt „Verarbeitungskapazität“ den gebräuchlicheren Begriff „schlussfolgerndes Denken“ als Äquivalent verwenden.

3 Aufbau des GkKT

3.1 Testkonstruktion

Im Rahmen des vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Projektes „Einstieg mit Erfolg“ an der Justus-Liebig-Universität Gießen wurden mehrere OSA entwickelt. Um schlussfolgerndes Denken zu erfassen, entwickelten wir den GkKT. Ziel der Testkonstruktion war einerseits, schlussfolgerndes Denken möglichst breit zu erfassen und andererseits sehr anwendungsorientiert einen Test mit möglichst geringer Bearbeitungszeit und hoher Akzeptanz zu entwickeln. Basierend auf umfangreichen Recherchen zu Intelligenztheorien und zu klassischen Aufgabentypen in Intelligenztests, wurden diverse Aufgabengruppen entwickelt und Pretests bei Studienanfängern unterzogen. Die Aufgabengruppen mit guten psychometrischen Eigenschaften stellten wir dann zum GkKT zusammen. Um die Breite der Erfassung des schlussfolgernden Denkens zu gewährleisten, entschieden wir uns, alle drei Inhaltsbereiche gemäß Jäger (1984) abzudecken. Zusätzlich war eine zielgruppengerechte Sprache in der Testadministration sowie der Testgestaltung zu bedenken. Dazu wurden Studierende des Bachelorstudiengangs Psychologie, die an einem Seminar zur Verfahrensentwicklung teilnahmen, in die Entwicklung der Aufgabengruppen mit einbezogen. Dadurch konnten wir eines der übergeordneten Ziele des Gesamtprojekts Einstieg mit Erfolg verwirklichen: Studierende als Akteure in die Projekte einzubeziehen. Dies hat den Vorteil, dass Projekthinhalte und damit auch im Projekt entwickelte Instrumente wie der GkKT maßgeblich von Zielgruppen nahen Personen mitgestaltet wurden, denn die Studienorientierungsphase lag bei den beteiligten Studierenden in der Regel nur zwei Jahre zurück.

3.2 Aufbau und Durchführung des GkKT

Der GkKT besteht aus 12 Aufgabengruppen, deren Administration je nach Anzahl an Items und Komplexität zwischen einer halben bis circa sechs Minuten dauert. Insgesamt dauert die Administration ca. 28.5 Minuten. Von den insgesamt 66 Items entfallen etwa gleich viele auf jede Inhaltskategorie (verbal, numerisch, figural). In Tabelle 1 sind die Aufgabengruppen mit der jeweiligen Anzahl an Items sowie der Zuordnung zu den Inhaltsbereichen gemäß Konstruktion aufgelistet.

Tabelle 1: Zuordnung der Aufgabengruppen zu den Inhaltsbereichen

Aufgabengruppe	Anzahl Items	Inhaltsbereich
Sprichwörter	2	v
Antonyme	4	v
Analogien	5	v
Syllogismen	3	v
Wortkette	5	v
Beziehungen	3	v
Verhältnis	6	n
Zahlengitter	8	n
Zahlenreihen	9	n
Figuren	7	f
Puzzle	7	f
Labyrinth	7	f

Anmerkungen: Zuordnung gemäß Intention bei der Testkonstruktion, v = verbal, n = numerisch, f = figural

Tabelle 2 gibt zu jeder Aufgabengruppe eine Beschreibung eines prototypischen Items wieder. Zur besseren Illustration der figuralen Items, sind in Abbildung 1 für die drei Aufgabengruppen Figuren, Puzzle und Labyrinth die Beispielitems, die in der Testinstruktion enthalten sind, zu finden.

Tabelle 2: Prototypische Itembeschreibungen

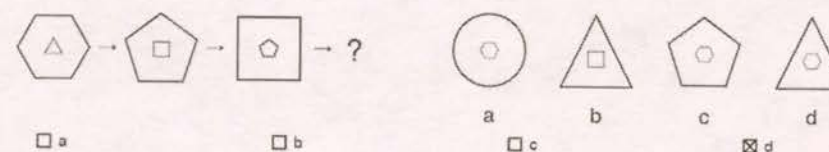
Aufgabengruppe	Itembeschreibungen
Sprichwörter	Es werden fünf Sprichwörter vorgegeben, von denen genau zwei die gleiche inhaltliche Bedeutung haben. Diese beiden sind durch Ankreuzen zu markieren.
Antonyme	Zu einem vorgegebenen Adjektiv muss aus einer Auswahl von fünf anderen Adjektiven dasjenige ausgewählt werden, das von der Bedeutung das Gegenteil des vorgegebenen ist.
Analogien	Ein Wortpaar wird vorgegeben, bei dem die beiden Worte in einer logischen Beziehung zu einander stehen. Aus einer Auswahl von fünf weiteren Wortpaaren muss das Wortpaar identifiziert werden, das eine ähnliche Beziehung der Worte zueinander aufweist.
Syllogismen	Es werden zwei Aussagen (Prämissen) vorgegeben, von denen angenommen werden soll, sie seien wahr. Aus einer Auswahl von vier möglichen

Aufgabengruppe	Itembeschreibungen
	Schlussfolgerungen muss diejenige ausgewählt werden, die sich folgerichtig aus den zwei Aussagen (Prämissen) ergibt.
Wortkette	Wortreihen bestehend aus drei Worten, die in einer regelbasierten Reihenfolge geordnet sind, müssen an der vorgegebenen Position in der Reihe um ein weiteres Wort ergänzt werden. Dazu muss eines von vier zur Auswahl stehenden Worten ausgewählt werden.
Beziehungen	Vorgegebene Wortpaare müssen dahingehend beurteilt werden, ob die zwei Worte eine ähnliche oder eine gegensätzliche Bedeutung haben oder weder noch.
Textrechnen	In wenigen Sätzen werden Informationen gegeben, die genutzt werden sollen, um mittels einer kurzen Rechnung (Grundrechenarten) die Antwort auf eine Frage zu ermitteln.
Zahlengitter	Zahlengitter (Matrizen) der Größe 3x3 sind an acht von neun Positionen mit Zahlen ausgefüllt. Es soll die Regel erkannt werden, nach der das Zahlengitter aufgebaut ist und die fehlende neunte Zahl in ein Freitextfeld eingetragen werden.
Zahlenreihen	Sieben Zahlen in einer Zahlenreihe sind vorgegeben. Es soll die Regel erkannt werden, nach der die Zahlenreihe aufgebaut ist und die Zahl, die an Stelle des „?“ in der Zahlenreihe stehen muss, soll in ein Freitextfeld eingetragen werden.
Figuren	Eine Reihe aus drei geometrischen Figuren muss um eine vierte Figur ergänzt werden. Dazu muss die Regel erkannt werden, nach welcher die Figurenreihe aufgebaut ist und aus einer Auswahl von vier Figuren die passende ausgesucht werden.
Puzzle	Es wird ein mehrteiliges Puzzle präsentiert, in dem ein Puzzleteil fehlt. Aus einer Auswahl von verschiedenen Puzzleteilen muss dasjenige ausgewählt werden, welches in die Lücke passt.
Labyrinth	Vorgegeben wird die Karte eines Labyrinths, die in gleichgroße Quadranten zerteilt ist. Am Rande der Karte sind ein Startpunkt und ein Zielpunkt markiert. Vom Startpunkt aus soll es einen durchgängigen Weg zum Zielpunkt geben. Dafür müssen manche Quadranten um 90 Grad nach links oder rechts gedreht werden, während andere in der Ausgangsstellung bleiben. Es soll für jeden der Quadranten angegeben werden, ob und ggf. wie er gedreht werden muss.

a)

In den folgenden Aufgaben werden Sie jeweils eine Figur sehen, die sich zweimal verändert. Ihre Aufgabe ist es, die Regel(n) der Veränderung zu erkennen und die richtige der vier vorgegebenen Figuren anstelle des Fragezeichens auszuwählen. Die Veränderung kann einer oder auch mehrerer Regeln folgen. Es gibt nur eine richtige Antwort, bitte kreuzen Sie diese an.

Beispielaufgabe:



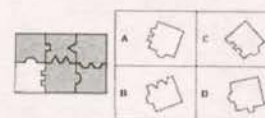
In diesem Beispiel ist Antwort d richtig, da die äußere Form der Figur nach jedem Schritt eine Ecke weniger hat, die innere Figur jedoch jeweils eine Ecke mehr.

b)

Bei den folgenden Aufgaben geht es darum, das fehlende Puzzlestück zu ergänzen. Eines der Puzzlestücke ergänzt die linke Abbildung. Eventuell muss das Puzzlestück gedreht werden, damit es passt. Es ist immer nur eine Lösung richtig.

Kreuzen Sie bitte den Buchstaben an, der das passende Puzzlestück kennzeichnet.

Beispielaufgabe:



Die Lösung lautet C

c)

Im folgenden Aufgabentyp geht es darum, den Weg vom Start zum Ziel zu finden.

Die Wegstücke können durch eine Links- oder Rechtsdrehung zu einem Weg verbunden werden. Überlegen Sie, wie die jeweiligen Wegstücke **gedreht** werden müssen damit Sie vom Startpunkt einen **durchgängigen Weg** zum Zielpunkt finden. Jedes Wegstück muss in den Weg mit eingebaut werden.

Sie können die Wegstücke um **90 Grad nach links** (gegen den Uhrzeigersinn) oder **rechts** (mit dem Uhrzeigersinn) **drehen** oder in **ihrer Position lassen**.

Beispiel:

Wie müssen die Wegstücke A, B, C, und D gedreht werden, um vom Start zum Ziel zu kommen?



Die richtige Antwort lautet: **A2** (Drehung nach rechts), **B0** (keine Drehung), **C1** (Drehung nach links), **D2** (Drehung nach rechts)

Abbildung 1: Beispielitems für die Aufgabengruppen mit figuralem Material:
a) Figuren, b) Puzzle und c) Labyrinth

3.3 Auswertung des GkKT

Der GkKT wird ausgewertet, indem für jedes korrekt gelöste Item genau ein Punkt vergeben wird. Summiert man die Itemscores aller Items mit verbalen, aller Items mit numerischem und aller Items mit figuralem Material jeweils auf, so erhält man die Kennwerte für die drei Skalen verbal, numerisch und figural. Summiert man über alle Items hinweg auf beziehungsweise bildet die Summe der drei Skalenwerte so erhält man den Gesamtscore im GkKT. Dieser liegt theoretisch im Intervall [0; 66].

4 Psychometrische Güte

Im Folgenden sollen die Befunde zur psychometrischen Güte dargestellt werden. Dazu wird zunächst die Eichstichprobe beschrieben, die die Datenbasis für die Untersuchungen zur Itemschwierigkeit und Itemtrennschärfe sowie zur Objektivität, Reliabilität und Validität lieferte. Zusätzlich werden Ergebnisse zur Untersuchung der Akzeptanz gegenüber dem GkKT und der Struktur des GkKT dargestellt. Diese Analysen erfolgten unter Verwendung der Software SPSS Version 21.0 (IBM Corp., 2012).

4.1 Eichstichprobe

Anhand von Paper-Pencil- und online-Testungen in verschiedenen Studiengängen (Chemie, Lebensmittelchemie, Materialwissenschaft, Psychologie und Soziale Arbeit) sowie anhand von Daten, die direkt online im OSA erhoben wurden, konnten wir eine Eichstichprobe mit $N = 803$ Personen (62.9 % Frauen) gewinnen. Mehr als die Hälfte (58.5 %) der Testteilnehmer gab an, zur Kategorie der 18- bis 19-Jährigen oder 20- bis 21-Jährigen zu gehören. Tabelle 3 zeigt die Altersverteilung in der Eichstichprobe.

Tabelle 3: Altersverteilung in der Eichstichprobe

Alterskategorie (Jahre)	<18	18-19	20-21	22-23	24-25	>25	k.A.
Anzahl Personen	62	287	183	64	43	143	21

Anmerkungen: k.A. = keine Angabe

4.2 Verteilung der Testwerte

Abbildung 2 zeigt die Verteilung des Gesamtscores des GkKT ($N = 803$). Gemäß Kolmogorov-Smirnov-Test ($Z = 1.041$, $p = .229$) liegt eine Normalverteilung vor. Der Test differenziert gut zwischen leistungsschwachen und -starken Teilnehmern ($M = 34.198$, $SD = 10.573$).

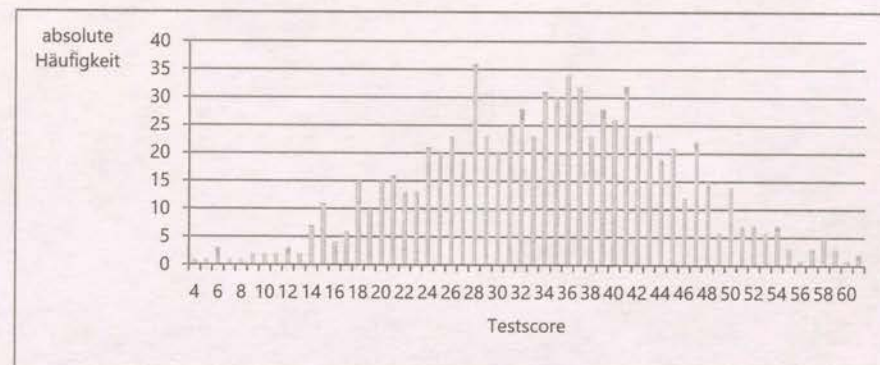


Abbildung 2: Verteilung des Gesamtscores ($N = 803$)

Tabelle 4 zeigt die deskriptiven Statistiken für den Gesamtscore im GkKT sowie die Scores in den drei Skalen. Die Scores für die verbale, numerische und figurale Skala verfehlen knapp die statistischen Anforderungen an eine Normalverteilung.

Tabelle 4: Deskriptive Statistiken zu den vier Kennwerten des GkKT

Kennwert	M (SD)	Schiefe (Standardfehler)	Kurtosis (Standardfehler)
Gesamtscore	34.198 (10.573)	-.102 (.086)	-.267 (.172)
Skala verbal	12.372 (3.674)	-.267 (.086)	-.048 (.172)
Skala numerisch	9.607 (4.916)	.249 (.086)	-.488 (.172)
Skala figural	12.211 (4.351)	-.115 (.086)	-.752 (.172)

Anmerkungen: M = Mittelwert SD = Standardabweichung; GkKT = Gießener kognitive Kompetenzen Test

4.3 Itemschwierigkeiten und Itemtrennschärfen

In Tabelle 5 sind die Schwierigkeiten und Trennschärfen aller 66 Items aufgeführt. Dabei gibt die Itemschwierigkeit wider, welcher Anteil der Stichprobe das Item korrekt gelöst hat. Die Itemtrennschärfe wiederum beschreibt, wie stark der Zusammenhang eines Items mit der jeweiligen Skala, der es zugeordnet wird, ist. Die Befunde zeigen, dass die innerhalb einer Aufgabengruppe weiter hinten stehenden Items wie gewünscht eine höhere Schwierigkeit aufweisen. Die Itemtrennschärfen spiegeln die oben genannten Befunde zu den internen Konsistenzen der drei Skalen wider. Sowohl die Itemschwierigkeiten als auch die Itemtrennschärfen liegen in erwarteten, mittleren Bereich.

Tabelle 5: Itemschwierigkeiten und Itemtrennschärfen aller Items des GkKT (N = 803)

Item	Itemschwierigkeit	Itemtrennschärfe
Sprichwörter_1	.350	.325
Sprichwörter_2	.630	.290
Antonyme_1	.880	.288
Antonyme_2	.885	.072
Antonyme_3	.851	.237
Antonyme_4	.301	.209
Analogien_1	.770	.346
Analogien_2	.777	.311

Item	Itemschwierigkeit	Itemtrennschärfe
Analogien_3	.547	.224
Analogien_4	.214	.227
Analogien_5	.168	.248
Syllogismen_1	.588	.240
Syllogismen_2	.391	.296
Syllogismen_3	.127	.159
Wortkette_1	.649	.329
Wortkette_2	.760	.340
Wortkette_3	.829	.409
Wortkette_4	.760	.417
Wortkette_5	.593	.364
Beziehungen_1	.640	.292
Beziehungen_2	.427	.253
Beziehungen_3	.233	.239
Zahlengitter_1	.791	.291
Zahlengitter_2	.568	.370
Zahlengitter_3	.687	.437
Zahlengitter_4	.670	.426
Zahlengitter_5	.482	.477
Zahlengitter_6	.432	.447
Zahlengitter_7	.162	.409
Zahlengitter_8	.162	.447
Zahlenreihen_1	.457	.384
Zahlenreihen_2	.519	.479
Zahlenreihen_3	.489	.506
Zahlenreihen_4	.467	.499
Zahlenreihen_5	.205	.466
Zahlenreihen_6	.196	.474
Zahlenreihen_7	.133	.358
Zahlenreihen_8	.153	.229
Zahlenreihen_9	.052	.280

Item	Itemschwierigkeit	Itemtrennschärfe
Textrechnen_1	.483	.425
Textrechnen_2	.453	.293
Textrechnen_3	.461	.379
Textrechnen_4	.492	.471
Textrechnen_5	.473	.470
Textrechnen_6	.407	.386
Figuren_1	.927	.287
Figuren_2	.690	.247
Figuren_3	.699	.357
Figuren_4	.714	.304
Figuren_5	.457	.299
Figuren_6	.441	.267
Figuren_7	.328	.218
Puzzle_1	.941	.163
Puzzle_2	.804	.299
Puzzle_3	.814	.191
Puzzle_4	.797	.264
Puzzle_5	.639	.371
Puzzle_6	.721	.396
Puzzle_7	.386	.391
Labyrinth_1	.447	.523
Labyrinth_2	.467	.507
Labyrinth_3	.381	.499
Labyrinth_4	.431	.540
Labyrinth_5	.567	.618
Labyrinth_6	.299	.533
Labyrinth_7	.262	.510

Anmerkungen: GkKT = Gießener kognitive Kompetenzen Test

4.4 Objektivität

Bei der Durchführung der online Version des GkKT erhalten die Teilnehmer ausführliche schriftliche Anweisungen. Die Bearbeitungszeiten für die einzelnen Aufgabengruppen sind sekundengenau festgelegt und werden online durch automatisiertes Weiterblättern umgesetzt. Im Detail stehen für die Instruktion und die Bearbeitung der verschiedenen Aufgaben aus dem verbalen Inhaltsbereich 5 Minuten und 45 Sekunden, für die Aufgaben aus dem numerischen Inhaltsbereich insgesamt 11 Minuten und 10 Sekunden und für die Aufgaben aus dem figuralen Inhaltsbereich sowie 11 Minuten und 20 Sekunden zur Verfügung. Die Anzahl an Aufgaben pro Inhaltsbereich ist dabei zwar fast gleich (vgl. Tabelle 1), jedoch erfordern die Aufgaben aus dem numerischen und dem figuralen Inhaltsbereich deutlich längere Instruktionstexte und Bearbeitungszeiten, woraus die zeitliche Imbalance resultiert. Die automatische Auswertung erfolgt algorithmusbasiert. Der individuelle Testwert wird anhand der in 5. dargestellten Normierung für alle Testteilnehmer nach dem gleichen Schema interpretiert. Zusammenfassend kann die Objektivität des GkKT als gewährleistet angesehen werden.

4.5 Reliabilität

Bei den Untersuchungen zur Reliabilität liegt der Fokus auf der internen Konsistenz des Gesamttests und der Skalen verbal, numerisch und figural. Für den Gesamtscore konnten wir eine sehr gute interne Konsistenz erzielen: Cronbachs $\alpha = .898$ ($N = 802$). Das spiegelt sich auch im Äquivalent McDonalds Omega wider: $\omega_t = .900$ ($N = 802$). Die drei Skalen erwiesen sich ebenfalls als gut bis sehr gut reliabel (vgl. Tabelle 6). Dass die verbale Skala vergleichsweise die geringste interne Konsistenz aufweist ist vermutlich der Konstruktion geschuldet. So ist sie aus sechs verschiedenen Aufgabengruppen zusammengesetzt, wohingegen die numerische und die figurale Skala aus je drei Aufgabengruppen zusammengesetzt sind. Die größere Vielfalt an Aufgabentypen geht erwartungsgemäß zu Lasten der Homogenität.

Tabelle 6: Reliabilität (interne Konsistenz) der drei Skalen ($N = 803$)

Skala	Anzahl Items	Alpha
verbal	22	.727
numerisch	23	.850
figural	21	.815

4.6 Validität

Da die im Folgenden berichteten Validerungsstudien teilweise nur an Subgruppe der Eichstichprobe durchgeführt wurden, variiert die Anzahl an Personen, auf deren

Daten die Validitätsbefunde beruhen. Für die einzelnen Befunde ist daher stets angegeben, um welchen Stichprobenumfang es sich handelt.

4.6.1 Konstruktvalidität

In ersten Studien zum Nachweis konvergenter Validität im Sinne der Konstruktvalidität untersuchten wir in verschiedenen Teilstichproben Zusammenhänge des GkKT mit insgesamt drei verschiedenen Intelligenztests. Dazu kam eine frühere Version des GkKT (Ulfert, Ott, Michaelis & Kersting, 2014) zum Einsatz. Diese Version umfasst lediglich 45 der 66 Items. Die hier im Artikel beschriebene aktuelle Version des GkKT mit 66 Items enthält alle Items der früheren Version und wurde zusätzlich um 21 Items erweitert, um eine weitgehend paritätische Aufteilung der Itemanzahl auf die drei Inhaltsbereiche zu gewährleisten. Die frühere Version korreliert überaus substantiell zu $r = .967$, $p < .001$ ($N = 803$) mit der aktuellen Version des GkKT, sodass davon auszugehen ist, dass die Befunde zur Konstruktvalidierung übertragbar sind. Tabelle 6 zeigt, dass sich sowohl mit dem Mehrfachwahl-Wortschatz-Intelligenztest MWT-B (Lehrl, 2005), als auch mit dem BEFKI GC-K (Schipolowski et al., 2013) sowie dem Wonderlic Test (Wonderlic, 1983) bedeutsame Zusammenhänge nachweisen lassen (vgl. Tabelle 7). Dabei ist der Zusammenhang mit dem Wonderlic Test signifikant höher als mit dem MWT-B ($Z = 5.262$, $p < .001$) und als mit dem BEFKI GC-K ($Z = 5.108$, $p < .001$). Dies spricht dafür, dass der GkKT wie intendiert schlussfolgerndes Denken relativ breit erfasst. Im Gegensatz dazu sind die beiden anderen Verfahren sehr spezifisch auf eine Dimension ausgereichtet: der BEFKI GC-K (wie unter 1.5 beschrieben) und auch der MWT-B fokussieren auf kristalline Intelligenz unter Verwendung verbalen Itemmaterials. Entsprechend fällt der Zusammenhang dieser Verfahren mit dem GkKT geringer aus als mit dem Wonderlic Test, welcher zur Erfassung der allgemeinen Intelligenz konstruiert wurde. Diese Befunde sprechen folglich für die Konstruktvalidität des GkKT.

Tabelle 7: Zusammenhänge des GkKT mit anderen Intelligenztests

Intelligenztests	MWT-B	BEFKI GC-K	Wonderlic
Korrelation mit GkKT	.416** (169)	.489** (330)	.766** (181)

Anmerkungen: GkKT = Gießener kognitive Kompetenzen Test; MWT-B = Mehrfachwahl-Wortschatz-Intelligenztest; BEFKI-GC-K = Kurzskala zur Messung kristalliner Intelligenz; Wonderlic = Wonderlic Test

4.6.2 Kriterienvalidität

Die Kriteriumsvalidität in Form des Zusammenhangs mit Schulnoten (retrograd) konnte ebenfalls gezeigt werden (vgl. Tabelle 8). Für den Gesamtscore im GkKT ergaben sich die stärksten und auch bedeutsamen Zusammenhänge mit der Abiturnote ($r = -.312$, $p < .001$, $N = 103$) sowie der Mathematiknote ($r = -.360$, $p < .01$,

$N = 299$) (einfache, nicht korrigierte Korrelationen). Dass die Zusammenhänge negativ sind, ergibt sich aus der Schulnotenskala (1 = „sehr gut“ bis 6 = „ungenügend“). Die Höhe der Korrelationen entspricht der Kriteriumsvalidität anderer Intelligenztest (Kersting, 2008; Liepmann et al., 2007), die deutlich zeitaufwändiger sind und unter kontrollierten Bedingungen erhoben wurden. Der Skalenwert für den verbalen Teil des GkKT erwies sich als hochgradig valide bei der Aufklärung der Varianz innerhalb der Abiturnote. Der Skalenwert für den numerischen Teil zeigte insbesondere bedeutsame Zusammenhänge mit der Mathematik- und der Physiknote. Der Skalenwert für den figuralen Teil konnte besonders zur Varianzaufklärung innerhalb der Mathematik- und Chemienote beitragen.

Tabelle 8: Kriteriumsvalidität des GkKT: Korrelationen mit Schulnoten

Prädiktor	Kriterien (Schulnoten)						
	Mathematik (299)	Deutsch (297)	Englisch (109)	Biologie (189)	Chemie (183)	Physik (174)	Abitur (103)
Testwert GkKT	-.360**	-.156**	.002	-.213**	-.261**	-.289**	-.312**
Faktor verbal	-.144**	-.183**	-.034	-.179*	-.012	-.101	-.326**
Faktor numerisch	-.317**	-.113	.028	-.202**	-.270**	-.318**	-.217**
Faktor figural	-.346**	-.091	-.003	-.119	-.264**	-.211**	-.134**

Anmerkungen: GkKT = Gießener kognitive Kompetenzen Test

Da Studierende durchaus verlässlich Globalurteile über ihre kognitiven Fähigkeiten abgeben können (Paulhus, Lysy & Yik, 1998), setzten wir bei einer Teilstichprobe diese Selbsteinschätzungsmaße mit dem Testergebnis im GkKT in Beziehung. Dazu baten wir die Teilnehmer vor der Intelligenztestung um eine Selbsteinschätzung ihrer Intelligenz. Bei der Teilstichprobe der Materialwissenschaftler und Chemiker ($N = 110$) zeigte sich ein bedeutsamer Zusammenhang zwischen dem Testergebnis und der Selbsteinschätzung hinsichtlich zahlengebundenem Denken ($r = -.238$, $p = .012$), nicht jedoch mit der Selbsteinschätzung hinsichtlich sprachgebundenem und räumlichem Denken ($r = .040$, $p = .681$ bzw. $r = -.067$, $p = .484$). Dass die Zusammenhänge negativ sind, ergibt sich aus dem Antwortformat bei den Selbsteinschätzungen (1 = „sehr hoch“ bis 6 = „sehr niedrig“).

Die konkurrente Kriteriumsvalidität sollte schließlich auch anhand der Zusammenhänge mit der aktuellen Durchschnittsnote im Studium aufgezeigt werden. Anhand einer Teilstichprobe vom Umfang $N = 142$ konnte ein erwartungsgemäßer, jedoch

nur marginal signifikanter Zusammenhang zwischen dem Testergebnis im GkKT und der Note in einer Klausur zur Einstiegsvorlesung Psychologie für Bildungswissenschaftler (Noten vom Prüfungsamt; kein Selbstbericht) festgestellt werden: $r = -.156$, $p = .064$. Auch hier ergibt sich das negative Vorzeichen aus der Codierung der Note (Schulnotenskala).

4.7 Akzeptanz

Als Nebengütekriterium wurde die Akzeptanz der Testteilnehmer gegenüber dem GkKT untersucht. Erfasst wurde die Akzeptanz gegenüber der früheren Version mit 45 Items (Ulfert et al., 2014) mit dem Fragebogen Akzept-L (Kersting, 2008), der speziell zur Erfassung der Akzeptanz gegenüber Leistungsverfahren konzipiert ist. Jeweils anhand mehrerer Items wurden die vier Dimensionen Kontrollierbarkeit, Messqualität, Augenscheinvalidität und Belastungsfreiheit anhand einer sechsstufigen Antwortskala (1 = „trifft nicht zu“ bis 6 = „trifft genau zu“) bewertet. Abschließend wurde eine Gesamtbeurteilung über den Test erfragt (Die Gesamtbeurteilung ist also kein Mittelwert der Einzeldimensionen, sondern ein eigenständiges Urteil.). Das Gesamturteil zur Akzeptanz des GkKT beträgt $M = 3.48$ ($SD = 1.07$), eine Bewertung im leicht positiven Bereich. Dieses Akzeptanzurteil korreliert bedeutsam mit dem Gesamtscore im GkKT: $r = -.410$, $p < .001$ ($N = 512$). Personen, die im GkKT bessere Leistungen erzielen, bewerten den Test besser. Der Befund passt zu vorherigen Befunden (Kersting, 2008). Die Bewertungen der einzelnen Dimensionen der Akzeptanz können Tabelle 9 entnommen werden. Da nicht alle Teilnehmer der Akzeptanzuntersuchung alle Items beantworteten, schwankt der Umfang der Teilstichprobe, die pro Dimension herangezogen werden kann. Vergleicht man die Bewertung der vier Dimensionen untereinander, so wird deutlich, dass die Kontrollierbarkeit die im Mittel nominell am besten bewertete Dimension ist. Das ist ein typischer Befund für Leistungstest und insbesondere Intelligenztests (Kersting, 2008). Vergleiche zur Akzeptanz anderer Leistungstests sind schwer zu treffen. Zwar berichtet Kersting (2008) die Akzeptanzwerte für andere Intelligenztests (BIS-4, d2, IST-2000-R, Raven und WIT-2), aber diese wurden in einer anderen Situation erhoben. Insbesondere handelte es sich um beaufsichtigte Testungen. Bemerkenswert ist, dass auf der Ebene der Einzeldimensionen der Akzeptanz die anderen Tests nominell besser bewertet werden, auf der Ebene der Gesamtbeurteilung aber der GkKT besser abschneidet als alle übrigen Tests in der Untersuchung von Kersting (2008), mit Ausnahme des WIT-2. Möglicherweise stellte die unbeaufsichtigte online-Testung eine Situation dar, die dazu führt, dass andere Akzeptanzgesichtspunkte bedeutsam werden, die vom Akzept-L nicht direkt, sondern nur indirekt in Form des Gesamturteils erfasst werden.

Tabelle 9: Akzeptanzurteil zum GkKT ($N = 418$ bis 512)

	Kontrollierbarkeit	Messqualität	Augenscheinvalidität	Belastungsfreiheit	Gesamtbeurteilung
M	4.201	3.250	2.375	3.672	3.480
(SD)	(1.045)	(.888)	(.887)	(1.057)	(1.071)

Anmerkungen: Die vier Dimensionen wurden auf einer sechsstufigen Skala (1 = „trifft nicht zu“ bis 6 = „trifft genau zu“) erhoben. Die Gesamtbeurteilung erfolgte auf der Schulnotenskala; GkKT = Gießer kognitive Kompetenzen Test; M = Mittelwert SD = Standardabweichung

4.8 Dimensionsanalysen

Die intendierte Faktorenstruktur konnte unter Verwendung des lavaan-Pakets (Rossee, 2012) in R Version 3.4.4 (R Core Team, 2018) nachgewiesen werden: So weist ein hierarchisches Modell mit einem übergeordneten g-Faktor und drei darunterliegenden Faktoren (für jeden Inhaltsbereich einer) insgesamt einen guten Modellfit auf: CFI = .711, SRMR = .051, RMSEA = .040 [.039; .042], $\chi^2/df = 2.316$. Wir berichten hier vier Fit-Indizes und folgen somit der Empfehlung von Hooper, Coughlan und Mullen (2008), eine Kombination verschiedener Fit-Indizes zu berichten. Gemäß Hooper et al. (2008) sowie Hu und Bentler (1999) erfüllt das Modell in Bezug auf die Indizes SRMR und RMSEA sowie den χ^2 -Wert die Anforderungen für einen guten Fit. Betrachtet man dagegen den CFI, so wird die konventionelle Cutoff-Grenze für einen guten Fit deutlich verfehlt. Folgt man den Empfehlungen von Iacobucci (2010) und Fan (2016), so sollte einem einzelnen Fit-Index – und insbesondere dem CFI – kein so großes Gewicht bei der Gesamtbeurteilung zugesprochen werden. Zudem gibt es Befunde dazu, dass der CFI besonders sensitiv für Missspezifikationen der Faktorladungen ist (Hu & Bentler, 1998). Das könnte darauf hindeuten, dass zwar die getestete hierarchische Struktur im Allgemeinen durchaus plausibel ist, dass jedoch die konkrete relative Höhe der Faktorladungen nicht ganz korrekt spezifiziert wurde. Dies sollte in einer weiteren Untersuchung exploriert werden. Die Zusammenhänge zwischen dem Testwert insgesamt und den drei Skalenwerten waren ebenfalls hoch signifikant, wie Tabelle 10 zu entnehmen ist.

Tabelle 10: Korrelation des Testwertes gesamt mit den drei Skalenwerten (N = 802 bzw. 803)

Subskala	1	2	3	4
Testwert gesamt	1			
Faktor verbal	.755**	1		
Faktor numerisch	.848**	.453**	1	
Faktor figural	.832**	.478**	.547**	1

Anmerkungen: ** = $p < .01$

5 Normierung und Feedback

Dem Screening-Charakter des Tests entsprechend verzichten wir bei der Ergebnismeldung zum GkKT auf eine feine Differenzierung und melden, basierend auf den Quartilgrenzen der Verteilung, lediglich drei Leistungskategorien zurück: unterdurchschnittlich (0 bis 27 Punkte), durchschnittlich (28 bis 42 Punkte) sowie überdurchschnittlich (43 bis 66 Punkte). Die Testrückmeldung im OSA erfolgte in zwei Schritten. Zunächst wird dem Teilnehmer rückgemeldet, wie viele Items er korrekt gelöst hat. Dieses individuelle Ergebnis wird dann in Relation gesetzt zur Verteilung der Testwerte innerhalb der oben beschriebenen Eichstichprobe. Konkret wird der Prozentrang, den der Teilnehmer erreicht hat, angegeben und eine verbale Einordnung des Testwertes in eine der drei Kategorien unterdurchschnittlich/durchschnittlich/überdurchschnittlich gemeinsam mit einer grafischen Repräsentation dargeboten.

6 Diskussion

6.1 Reliabler und valider Test

Die berichteten Befunde weisen den GkKT als reliablen und validen Test zur Erfassung des schlussfolgernden Denkens im OSA-Kontext aus. So ergaben sich gute bis sehr gute Kennwerte für die interne Konsistenz für den Gesamttest sowie die verbale, numerische und figurale Skala. Zudem konnten die Konstrukt- und Kriterienvalidität nachgewiesen werden.

6.2 Gute Itemkennwerte

Die Analysen auf Itemebene ergaben, dass die 66 Items eine gute Streuung der Itemschwierigkeiten (wenige leichte Items, viele Items mit mittlerer Schwierigkeit

sowie wenige schwere Items) aufweisen. Dabei steigt innerhalb der Aufgabengruppen die Itemschwierigkeit von den ersten zu den letzten Items an. Nahezu alle Items weisen die gewünschten hohen Trennschärfen auf.

6.3 Gute Akzeptanz

Der GkKT wurde von der untersuchten Stichprobe in Bezug auf die Akzeptanz insgesamt positiv beurteilt. Von den untersuchten Dimensionen der Akzeptanz erwies sich besonders die Kontrollierbarkeit als gut bewertet. Das bedeutet, die getesteten Personen hatten den Eindruck, während der Testung gut informiert zu sein, wie die Testung abläuft, was ihre Aufgabe ist und wie das Testergebnis zu Stande kommt.

6.4 Auswertung des GkKT

Wie unter 3.3 beschrieben, erhält man beim GkKT für jedes korrekt gelöste Item einen Punkt. Hier sollte zukünftig untersucht werden, ob das Ergebnis im GkKT noch höhere Validität aufweisen kann, wenn bei der Bildung des Gesamtscores die Items nach ihrer psychometrischen Qualität gewichtet werden.

6.5 Weitere Validierungsstudien und Untersuchungen zur Akzeptanz wünschenswert

Die Konstruktvalidierung anhand der Zusammenhänge mit anderen Intelligenztests kann aktuell nur für die frühere GkKT-Version mit 45 Items berichtet werden. Zwar ist der Zusammenhang der früheren Version mit der hier vorgestellten Version mit 66 Items sehr eng, was erlaubt, diese Zusammenhänge als Indizien für die Konstruktvalidität des GkKT zu verstehen, dennoch sollten diese Untersuchungen zur Konstruktvalidität auch für die Version mit 66 Items erfolgen. Gleiches gilt für die Befunde zur Akzeptanz. Zusätzlich sind Untersuchungen zur prädiktiven Validität des GkKT in Bezug auf Studiennoten in Arbeit.

6.6 Unbeaufsichtigtes Testen

Im Rahmen von OSA stellt sich grundsätzlich die Frage nach der Belastbarkeit der Testergebnisse und damit auch die Frage nach der Generalisierbarkeit der Interpretationen. Ein wichtiger Aspekt ist, dass die OSA unbeaufsichtigt und anonym bearbeitet werden. Somit ist nicht sichergestellt, dass der Test ohne größere Störungen kontinuierlich bearbeitet wurde. Zudem ist es dem Teilnehmer überlassen, wie gewissenhaft er die Aufgaben bearbeitet. Wir empfehlen verschiedene Maßnahmen, um diesem Problem zu begegnen: Zu Beginn der Testung sollten die Teilnehmer eindringlich darauf hingewiesen werden, dass der Test etwa 30 Minuten dauern wird und sie sicherstellen sollen, dass sie in dieser Zeit ungestört durchgängig den Test bearbeiten. Zudem sollte darauf hingewiesen werden, dass keinerlei Hilfsmittel zur Testbearbeitung zugelassen sind.

Darüber hinaus sollte den Teilnehmern direkt bevor sie Feedback bekommen, bewusst gemacht werden, dass die Art und Weise der Bearbeitung ihr Ergebnis beeinflusst haben könnte. Beim Einsatz des GkKT im OSA der Justus-Liebig-Universität Gießen wird das Feedback eingeleitet mit: „Bitte denken Sie beim Lesen der nun folgenden Rückmeldungen immer an Folgendes: Da das Ergebnis des Self-Assessments außer von Ihren tatsächlichen Fähigkeiten von vielen weiteren Faktoren abhängt, wie beispielsweise Ihrer Stimmungslage, Ihrer Motivation oder Ihrer Konzentrationsfähigkeit zum Zeitpunkt der Bearbeitung, sollten Sie sich bei der Interpretation Ihres Ergebnisses auch fragen: Wurden Sie während des Self-Assessments von irgendetwas um Sie herum abgelenkt? Haben Sie sich ausschließlich auf das Self-Assessment konzentriert oder haben Sie Ihre Gedanken auch mal schweifen lassen? Wie ernst haben Sie das Self-Assessment genommen? Haben Sie bei manchen Aufgaben früher aufgegeben, weil Sie keine Lust mehr hatten? Diese Fragen sollen Ihnen beispielhaft verdeutlichen, dass unser Feedback an Sie durch Störfaktoren verfälscht sein kann und Sie unter anderen Umständen möglicherweise ein (noch) besseres Ergebnis hätten erzielen können.“ Zudem besteht in diesem Setting die Gefahr, die „Diagnose“ über zu bewerten und aufgrund der automatisierten Befundung besteht keine Möglichkeit, Interpretationsmissverständnisse aufzufangen. Diesen Risiken sollte durch die überaus sorgfältige und zurückhaltende Gestaltung der Feedbacks Rechnung getragen werden. Zudem empfiehlt sich ein abschließender Verweis auf weitere Beratungsangebote, wie beispielsweise persönliche Gespräche mit Studienberatern.

Schließlich können im Datensatz Plausibilitätsprüfungen vorgenommen werden. So kann man die Auswertung unterlassen, wenn bei einzelnen Fällen bestimmte kritische Cutoff-Werte in den Meta-Daten über- oder unterschritten werden (beispielsweise die Bearbeitungszeit unterschritten wurde oder ein unplausibles Antwortmuster vorliegt). Diese Plausibilitätsprüfungen können jederzeit von der Person, die die Daten auswertet, individuell vorgenommen werden. Abhängig von dem Tool, welches zur online-Umsetzung verwendet wird, kann dies auch automatisiert geschehen. In dem OSA, in dem der GkKT aktuell eingesetzt wird, ist die automatisierte Plausibilitätsprüfung noch nicht implementiert.

6.7 Fazit

Mit dem GkKT steht ein reliabler und valider Screening-Test zur Erfassung des schlussfolgernden Denkens für den Einsatz in OSA zur Verfügung. Der Test ist kurz (weniger als 30 Minuten), findet eine ausreichende Akzeptanz und wird auf Anfrage mit einer freien Lizenz angeboten. Damit werden bisherige Schwierigkeiten beim Einsatz von Intelligenztests in OSA überwunden. Sollte der Test sich auch an anderen

Hochschulen etablieren und häufig eingesetzt werden, könnten auf der derart resultierenden umfassenden Datenbasis sowohl differenziertere Normen (nicht nur eine Dreier-Einteilung) als auch Mehrfachnormen (für spezifische Studiengänge) berechnet werden.

Literatur

- Brunswick, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley and Los Angeles: The University of California Press.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin*, 40 (3), 153-193.
- Fan, J. (2016). The role of thinking styles in career decision-making self-efficacy among university students. *Thinking Skills and Creativity*, 20, 63-73.
- Franzke, B. (2010). Vermittlung von Berufsbildern – Wirkung und Relevanz von Rollenmustern und Geschlechterstereotypen bei der Beratung junger Menschen. Konzeption eines Gendertrainings. Verfügbar unter: https://www.professor-franzke.de/pdf/Franzke_Vermittlung_von_Berufsbildern_Gendertraining_07_2010.pdf [12.10.2018].
- Hasenberg, S. & Schmidt-Atzert, L. (2014). Internetbasierte Selbsttests zur Studienorientierung. *Beiträge zur Hochschulforschung*, 36 (1), 8-28.
- Heublein, U., Ebert, J., Hutzsch, C., Isleib, S., König, R., Richter, J. & Woisch, A. (2017). Zwischen Studiererwartungen und Studienwirklichkeit: Ursachen des Studienabbruchs, beruflicher Verbleib der Studienabbrecherinnen und Studienabbrecher und Entwicklung der Studienabbruchquote an deutschen Hochschulen. Hannover: Deutsches Zentrum für Hochschul- und Wissenschaftsforschung.
- Heukamp, V., Putz, D., Milbradt, A. & Hornke, L. (2009). Internetbasierte Self-Assessments zur Unterstützung der Studienentscheidung. *Zeitschrift für Beratung und Studium*, 4 (41), 2-8.
- Heydasch, T., Renner, K. H., Haubrich, J., Hilbig, B. & Zettler, I. (2013). A free web-based intelligence test: the hagen matrices test (hmt). unpublished manuscript. Hagen: University of Hagen.
- Hochschule der Bundesagentur für Arbeit und der Fachhochschule Nordwestschweiz (2018). OSA-Portal: Übersicht deutschsprachiger Online-Self-Assessments zur Studienorientierung. Verfügbar unter: <http://www.osa-portal.de/index.php> [03.10.2018].
- Hooper, D., Coughlan, J. & Mullen, M. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *The Electronic Journal of Business Research Methods*, 6 (1), 53-60.
- Hu, L. & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3 (4), 424-453.
- Hu, L. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6 (1), 1-55.
- Iacobucci, D. (2010). Structural equations modeling: Fit Indices, sample size, and advanced topics. *Journal of Consumer Psychology*, 20 (1), 90-98.
- IBM Corp. (2012). Released 2012. IBM SPSS Statistics for Windows, Version 21.0 [Computer software]. Armonk, NY: IBM Corp.
- Isleib, S. (2015). Neue Theorieströmungen zum Studienabbruch: Herkunft, Genese und Potenziale für die Studienabbruchs- und Hochschulforschung. Kassel: Deutsches Zentrum für Hochschul- und Wissenschaftsforschung.
- Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychologische Rundschau*, 35 (1), 21-35.
- Jäger, A. O., Süß, H.-M. & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test*. Göttingen: Hogrefe.
- Kersting, M. (2008). Zur Akzeptanz von Intelligenz- und Leistungstests. *Report Psychologie*, 33 (9), 420-433.
- Kersting, M., Althoff, K. & Jäger, A. O. (2008). WIT-2. Der Wilde-Intelligenztest: Verfahrenshinweise. Göttingen: Hogrefe.

- Lehrl, S. (2005). Mehrfachwahl-Wortschatz-Intelligenztest MWT-B (5. Auflage). Balingen: Spitta Verlag.
- Liepmann, D., Beauducel, A., Brocke, B. & Amthauer, R. (2007). Intelligenz-Struktur-Test 2000 R (I-S-T 2000 R): 2. erweiterte und überarbeitete Auflage. Göttingen: Hogrefe.
- McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, 37 (1), 1-10.
- Paulhus, D. L., Lysy, D. C. & Yik, M.S.M. (1998). Self-report measures of intelligence: Are they useful as proxy IQ tests? *Journal of Personality*, 66 (4), 525-554.
- R Core Team. (2018). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Verfügbar unter: <https://www.R-project.org> [27.09.2018].
- Richardson, M., Abraham, C. & Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychological Bulletin*, 138 (2), 353-387.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48 (2), 1-36.
- Salgado, J. F., Anderson, N., Moscoso, S., Bertuà, C., Fruyt, F. de & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology*, 88 (6), 1068-1081.
- Schipolowski, S., Wilhelm, O., Schroeders, U., Kovaleva, A., Kemper, C. J. & Rammstedt, B. (2013). BEFKI GC-K: Eine Kurzskaala zur Messung kristalliner Intelligenz. *Methoden, Daten, Analysen (mda)*, 7 (2), 153-181.
- Schmidt, F. L. & Hunter, J. E. (1998). The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings. *Psychological Bulletin*, 124 (2), 262-274.
- Schmidt, F. L. & Hunter, J. E. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, 86 (1), 162-173.
- Schmidt, F. L., Shaffer, J. A. & Oh, I.-S. (2008). Increased accuracy of range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology*, 61 (4), 827-868.
- Ulfert, A.-S., Ott, M., Bothe, P. S. & Kersting, M. (2017). Gießener kognitiver Kompetenztest (GkKT). Verfahrenshinweise. Version 2.0. Unveröffentlichtes Manuskript. Gießen: Justus-Liebig-Universität Gießen.
- Ulfert, A.-S., Ott, M., Michaelis, L. & Kersting, M. (2014). Gießener kognitiver Kompetenztest (GkKT). Verfahrenshinweise. Version 1.0. Unveröffentlichtes Manuskript. Gießen: Justus-Liebig-Universität Gießen.
- Whitmore, J. R. (1980). Giftedness, conflict, and underachievement. Boston: Allyn and Bacon.
- Wissenschaftsrat (2004). Empfehlungen zur Reform des Hochschulzugangs. Berlin.
- Wonderlic, E. F. (1983). Wonderlic personnel test manual. Northfield, IL: E.F. Wonderlic & Assoc.

Autor*innen

M.Sc. Pascale Stephanie Petri

Dipl.-Psych. Vera Weingardt

Prof. Dr. Martin Kersting

Justus-Liebig-Universität Gießen, Institut für Psychologische Diagnostik,

Projekt "Einstieg mit Erfolg" - Self-Assessment

Korrespondenz an: pascale.s.petri@psychol.uni-giessen.de

Empirische Pädagogik

33. Jahrgang 2019

3. Heft