Die DGP Testverfahren - Ein kurzer Rückblick und eine aktuelle Studie zur Konstrukt- und Kriteriumsvalidität des BIS-r-DGP Tests

Prof. Dr. Martin Kersting

1. Die DGP Testverfahren - Ein kurzer Rückblick

Die Deutsche Gesellschaft für Personalwesen (DGP) steht für Kompetenz in zahlreichen Bereichen der Personalpsychologie, von der Personalauswahl bis hin zur Personalentwicklung.
Der vorliegende Beitrag thematisiert nur einen Aspekt aus dem umfassenden Spektrum der DGP, die psychometrischen Tests – ein Gebiet, auf dem die DGP bundesweit Standards gesetzt hat und bis heute setzt. Neben Eignungsinterviews und Assessment-Centern (die DGP war eine der ersten Organisationen, die Assessment-Center in Deutschland durchgeführt haben) hat die DGP sich insbesondere um die Testdiagnostik verdient gemacht - 60 Jahre DGP, das bedeutet 60 Jahre Engagement für und Fortschritt im Bereich der Eignungstests. Zu Recht hat Prof. Irle die DGP als „Brutstätte quantitativer Testdiagnostik“ bezeichnet. Zahlreiche renommierte Wissenschaftler haben sich im Auftrag der DGP um die Testdiagnostik verdient gemacht, zu nennen sind u. a. J. v. Allesch, K. Althoff, R. Amthauer, H. Brandstätter, E. Fürntratt, S. Greif, A. O. Jäger, M. Irle, E. Todt und K. Wilde. Standardverfahren der Intelligenzdiagnostik wie der Intelligenz-Struktur-Test (IST) und der Wilde-Intelligenztest (WIT) wurden in der DGP entwickelt, diese Verfahren wiederum waren Grundlage zahlreicher anderer Entwicklungen wie z. B. dem Berliner-Intelligenzstrukturtest (BIS-4). Es wäre interessant zu untersuchen, inwieweit noch heute in dem mittlerweile fast unüberschaubaren und heterogenen Testmarkt die Elemente dieser „Ursuppe“ der Leistungstests in der Bundesrepublik wiederzuerkennen sind. Wer sich allein die drei zuletzt genannten Intelligenztests anschaut, erkennt rasch die sich aus den gemeinsamen Wurzeln ergebende hohe Verwandtschaft. Aber auch auf anderen Testgebieten hat die DGP die Fundamente gelegt, etwa mit dem Berufsinteressentest (BIT) oder dem Differentiellen Interessen-Test (DIT). Dass diese bislang genannten Verfahren einen hohen Bekanntheitsgrad haben liegt u. a. daran, dass sie bei einem Testverlag publiziert wurden. Dies ist aber eher untypisch für die DGP, die ihren Kunden eine exklusive Nutzung der Tests bei gleichzeitig maximaler Vertraulichkeit der Testmaterialien bietet und deren Verfahren daher zumeist so genannte „confidential tests“ sind. Derartige Verfahren werden ungeachtet ihrer Bedeutung zwangsläufig weniger bekannt, selbst wenn es sich um innovative und wegweisende Verfahren handelt, wie die DGP Tests zum büropraktischen Arbeiten (z. B. die so genannte Postaufgabe), semantisch in

Form einer Arbeitsprobe eingekleidete Intelligenztests (z. B. die so genannte „Textanalyse“ sowie die Aufgabe „Tabellen und Statistiken“) oder Kenntnistests zu verschiedenen Wissensdomänen sowie Tests zum technisch-mechanischen Verständnis. Auch der im Folgenden thematisierte „BIS-r-DGP“ Test ist in diesem Kontext zu nennen. Dieser Test ermöglicht eine ökonomische Erfassung der Dimensionen des Berliner Intelligenzstrukturmodells und ist wesentlich umfassender kriteriumsvalidiert als der BIS-4 Test, der „Originaltest“ zum Berliner Intelligenzstrukturmodell.

Nicht nur zu den zahlreichen hochwertigen Testkonstruktionen, sondern auch zu den empirischen Studien zur Evaluation muss man der DGP anlässlich des 60-jährigen Jubiläums gratulieren: Keine Institution im deutschsprachigen Raum hat mehr empirische Studien zur Kriteriumsvalidität von Personalauswahlverfahren durchgeführt und publiziert als die DGP, hinzu kommen etliche weitere empirische Studien zu spezifischen Fragestellungen der Eignungsdiagnostik, wie z. B. zur Fairness von Testverfahren oder zu Anforderungsanalysen. Schließlich ist das fach- und gesellschaftspolitische Engagement der DGP auf dem Gebiet der Eignungsdiagnostik zu würdigen, etwa der Beitrag der DGP zur Entwicklung der DIN 33430 (DIN, 2002) sowie aktuell zur Entwicklung der entsprechenden ISO Norm.

2. Zur Bedeutung und Messung der Intelligenz

Von Anfang an hat die DGP die Bedeutung der Intelligenz für den Ausbildungs- und Berufserfolg erkannt und mit der Entwicklung von Tests wie dem oben genannten BIS-r-DGP Test, dem IST und dem WIT maßgeblich zur Erforschung und praktischen Nutzung der Intelligenzmessung beigetragen. Die DGP hat auch dann an dem Konstrukt Intelligenz festgehalten, als dieses in den 1970er Jahren gesellschaftlich verpönt war. Mittlerweile ist die Bedeutung der Intelligenz unumstritten, nach der Sichtung metaanalytischer Studien resümieren Schmidt und Hunter (1998), dass aufgrund von Intelligenztests gewonnene Aussagen die höchste Validität bei der Vorhersage zukünftiger Leistungen erzielen, und zwar sowohl bei der Vorhersage von Ausbildungs- als auch bei der Vorhersage von Berufsleistungen. Hunter und Hunter (1984) berichten entsprechende gemittelte Koeffizienten von .54 für den Ausbildungs- und .45 für den Berufserfolg. Salgado, Anderson, Moscoso, Bertua und De Fruyt (2003) berechnen für die Intelligenz anhand von europäischen Datensätzen eine Kriteriumsvalidität in Höhe von .62. Eine Metaanalyse der deutschen Studien zur Kriteriumsvalidität von Intelligenztests wurde

von Hülsheger, Maier und Stumpp (2007) berechnet. Die Autoren kommen auf einen Vorhersagewert von .47 für Ausbildungserfolg und .53 für Berufserfolg. Nach der neuesten metaanalytischen Übersicht von Kramer (2009) ergeben sich für die Intelligenz Koeffizienten von .62 für die Vorhersage von Arbeitsleistungen, .59 für die Vorhersage von Lernleistungen, .33 für die Vorhersage des Einkommens und .31 für die Vorhersage von beruflichen Entwicklungen. Als Fazit einer bald einhundertjährigen Forschung zu diesem Thema mit Daten von mehreren Zehntausend Personen kann eindeutig festgehalten werden: Mit keinem anderen Verfahren lässt sich der Erfolg bei kognitiv geprägten Lebensaufgaben wie Schule, Ausbildung, Studium und Beruf so gut vorhersagen wie mit Intelligenztests. Die Variabilität der Vorhersageleistung über verschiedene Situationen und vor allem über verschiedene Berufe hinweg ist gering. Dies bedeutet, dass Intelligenztests nicht nur für spezifische (Berufs-)Gruppen, sondern generell valide Schlussfolgerungen erlauben (Schmidt et al., 1993; Salgado, Anderson, Moscoso, Bertua, De Fruyt & Rolland, 2003).

3. Zur praktischen Bedeutung der Konstruktvalidierung

Wenn die Intelligenz, wie im vorherigen Abschnitt referiert, ein so bedeutsames Konstrukt ist, ist es gut zu wissen, dass ein Test nicht nur behauptet, Intelligenz zu messen, sondern es auch wirklich tut. Dies prüft man mit einer Validierungsstrategie, deren Sinn sich den Praktikern auf den ersten Blick vielleicht nicht erschließt, mit der sogenannten Konstruktvalidierung. Bei der Konstruktvalidierung geht es darum, zu prüfen, ob der Test das misst, was er zu messen beabsichtigt und ob die aufgrund der Testergebnisse getroffenen Interpretationen theoriekonform sind. Man versucht, “die Beziehungen zwischen Testverhalten und theoretischen Begriffen (den Konstrukten)” zu klären (Jäger, 1986, S. 203). Das klingt zunächst sehr akademisch; den Praktiker interessieren weniger Theorien, sondern die Vorhersagekraft, also die Kriteriumsvalidität. Bei der Kriteriumsvalidierung prüft man, ob der Testwert die Bestimmung von Nicht-Testverhalten erlaubt, beispielsweise die Vorhersage des Ausbildungs- oder Berufserfolgs. Aber auch hier zeigt sich, dass nichts praktischer ist als eine gute Theorie. Über die Konstruktvalidität kann man beispielsweise Rückschlüsse auf die Treffsicherheit eines Verfahrens, also die Kriteriumsvalidität, ziehen. Dies funktioniert so: Wenn in mehreren hundert Studien mit den Daten von mehreren Zehntausend Personen gezeigt wurde, dass Intelligenztests allgemein eine treffsichere Vorhersage des Erfolgs bei kognitiv geprägten Lebensaufgaben erlauben (siehe oben), kann ein einzelner Test unter bestimmten Umständen

bereits dann als kriteriumsvalide (also treffsicher) gelten, wenn er nachweislich Intelligenz misst, selbst wenn für diesen Test selbst keine Bewährungskontrollen in Bezug auf die Vorhersage von Kriterien (z. B. Ausbildungs- oder Berufserfolg) durchgeführt wurden. In der DIN 33430 (DIN, 2002) wird explizit auf die Möglichkeit eingegangen, Gültigkeitsbelege, die in anderen Studien erbracht wurden, auf ein anderes Verfahren zu übertragen (Validitätsgeneralisierung).

Dies klingt nun so, als wäre die Konstruktvalidierung lediglich ein Umweg – und manchmal eine Abkürzung – zur eigentlich bedeutsamen Kriteriumsvalidität. Dabei gilt - genau umgekehrt - das Primat der Konstruktvalidität. Ohne gültige Konstruktannahmen ist die Kriteriumsvalidität wenig nützlich. Erst die Konstruktannahmen stellen einen Rahmen zur Integration verschiedener Befunde (zu einem Konstrukt, nicht etwa nur zu einem Verfahren) dar und sind somit das Fundament eines kumulativen Erkenntnisgewinns. Eine Beschränkung der Berufseignungsdiagnostik auf Kriteriumsvalidierungen ist „blinde Technologie, solange die Zusammenhänge zwischen Test- und Kriteriumsverhalten nicht psychologisch erhellt, und das heißt theoretisch auf den Begriff gebracht sind“ (Jäger, 1986, S. 284). Ohne eine theoretische Einbettung sind Kriteriumsvaliditäten ggf. sogar irreführend, denken wir nur an die signifikante Korrelation zwischen der Anzahl an Störchen und der Anzahl an Geburten. Diese Korrelation beträgt R=.49 (Höfer, Przyembel, Verleger, 2004), dennoch ist die Beeinflussung der Anzahl an Störchen keine sinnvolle Grundlage der Geburtenkontrolle: Die Korrelation ist auf eine gemeinsame Drittvariable zurückzuführen; die zunehmende Industrialisierung beeinflusst sowohl die Anzahl an Störchen als auch die Anzahl an Geburten. Auch die bedeutsame Korrelation zwischen Schuhgröße und Einkommen sollten niemanden verleiten, nur Personen mit großen Füßen auszuwählen: Frauen haben durchschnittlich kleinere Füße und werden durchschnittlich schlechter bezahlt. Die eigentlich interessanten Zusammenhänge bleiben unberührt, wenn man lediglich (signifikante) Zusammenhangsmuster („miscellaneous correlations“, Cronbach, 1989, S. 155) zwischen Prädiktoren und Kriterien betrachtet. Erst eine theoretische Einordnung dieses Zusammenhangs schöpft den Nutzen von Studien zur Kriteriumsvalidität aus. Für diese Einordnung ist es unabdingbar notwendig zu wissen, was gemessen wird. Entsprechend waren die ersten Arbeiten zur Konstruktvalidierung bahnbrechend für die Entwicklung der Psychometrie, der Fachaufsatz „Convergent and discriminant validation by the multitrait-multimethod matrix“ von Campbell und Fiske (1959) ist die am häufigsten zitierte Arbeit, die jemals im Psychological Bulletin erschienen ist, gefolgt von dem Aufsatz „Construct validity in psychologi-

cal tests“ (Cronbach & Meehl, 1955).

Die fundamentale Bedeutung von Konstruktannahmen für die Kriteriumsvalidierung wird im Folgenden für fünf Bereiche beispielhaft erläutert.

(1) Um die Validität eines Verfahrens erschließen zu können, bedarf es Annahmen darüber, was der Prädiktor misst. Messick (1988, S. 37) hat dies anhand von Beispielen verdeutlicht. So ergibt sich z.B. eine unterschiedliche Bewertung des nachweislichen Prädiktor-Kriteriums-Zusammenhangs, je nachdem, ob man ein und denselben Testwert als Ausprägung auf der Dimension „Flexibilität“ oder als Ausprägung der Dimension „Zerstreutheit“ interpretiert.

(2) Auch die Frage, ob man die Kriterien als solche akzeptiert, kann erst aufgrund der Interpretation der gemessenen Leistungsdimension entschieden werden. Eine unreflektierte Fixierung auf Prädiktor-Kriteriums-Zusammenhänge ist reaktionär. Während es sich bei den meisten Prädiktoren um psychologisch definierte Konstrukte handelt, stellen Außenkriterien sozial definierte Konstrukte dar (Wiggins, 1973). Diese sind starken Veränderungen (z.B. den jeweiligen Werten und Normen sowie dem technologischen Wandel) ausgesetzt. Ein „guter“ Polizist war beispielsweise früher eine Respektperson, die man nicht ohne weiteres angesprochen hat und ist heute als „Bürger in Uniform“ Ansprechpartner – entsprechend ändern sich die Kriterien als Indikatoren für den Erfolg eines Polizisten. Eine allein an – zwangsläufig zeitlich zurückliegenden - Studien zur prädiktiven Kriteriumsvalidität orientierte Berufseignungsdiagnostik bezieht sich auf Verfahren, die zeitlich gesehen *vor* einem möglichen Wandel der sozial definierten Kriterien oder vor einem möglichen technologischen Wandel zur Prognose eingesetzt wurden. Eine theorielose Berufseignungsdiagnostik ist also stets rückwärtsorientiert.

(3) Konstruktannahmen ermöglichen es, Prädiktoren und Kriterien systematisch aufeinander zu beziehen. Komplexe Kriterien wie der Berufserfolg können theoriegeleitet in einzelne Facetten zerlegt werden, um dann zu entscheiden, welche Facette durch welches Verfahren/welchen Prädiktor vorhersagbar ist (siehe z.B. das Symmetrie-Prinzip, Wittmann, 1988).

(4) Sofern differentielle Kriteriumsvaliditäten für einzelne Gruppen auftreten, ein Test also beispielsweise bei Männern eine höhere Kriteriumsvalidität aufweist als bei Frauen, kann ebenfalls nur vor dem Hintergrund theoretischer Annahmen entschieden werden, ob es sich um eine wünschenswerte Differenzierung oder um einen unerwünschten „bias" handelt.

(5) Daten zur Kriteriumsvalidierung wären schließlich von zweifelhaftem Wert, wenn ihr Geltungsanspruch sich ausschließlich auf die den Daten zugrunde liegende Situation beziehen würde. Der gängigen Interpretation von Kriteriumsvaliditäten liegen zumeist Annahmen zur zeitlichen Stabilität des prognostizierten und zu prognostizierenden Merkmals sowie Annahmen zur Situationsinvarianz – und somit Konstruktannahmen – zugrunde. Ohne Konstruktannahmen wären Generalisierungen von Kriteriumsvaliditäten und somit Absicherungen gegen bestimmte Arten von Stichproben- und Messfehler (siehe z.B. Schmidt, 1992) zumindest deutlich erschwert.

Aber auch aus zahlreichen anderen Gründen ist es überaus bedeutsam zu erfahren, was ein Test überhaupt genau misst. Neben der Möglichkeit der Validitätsgeneralisierung ist die Konstruktvalidität von zentraler diagnostischer Bedeutung u. a. für die Konstruktion und Auswahl diagnostischer Verfahren, für die zielgerichtete Kombination verschiedener Verfahren, für die Bewertung von empirischen Befunden, für die Interpretation der Befunde, für die Ableitung von Trainings- und Fördermaßnahmen sowie für die Theorieentwicklung (z. B. für die Entwicklung einer Theorie des Berufserfolgs).

Für die Konstruktvalidierung können zahlreiche Methoden genutzt werden, die zum größten Teil bereits bei Cronbach und Meehl (1955) genannt wurden. Man kann beispielsweise Gruppen untersuchen, die sich hinsichtlich des Konstrukts eindeutig unterscheiden sollten. Will man, so das Beispiel von Cronbach und Meehl (1955), mit einem Test religiöse Überzeugungen messen, sollten aktive Mitglieder einer Kirchengemeinde andere Ergebnisse erzielen als Personen, die keinen Bezug zu einer Glaubensgemeinschaft haben usw. Eine andere gängige Methode der Konstruktvalidierung besteht darin zu prüfen, ob die aufgrund der Theorie postulierte Struktur der Daten sich tatsächlich in den Daten widerspiegelt. Kersting und Beauducel (1997) haben beispielsweise aufgezeigt, dass sich in den Daten, die mit dem BIS-r-DGP Test gewonnen wurden, die vier Operations- und die drei Inhaltsklassen des Berliner Intelligenzstrukturmodells nachweisen lassen. In der im Folgenden dargestellten Studie wurde zur Konstruktvalidierung der Ansatz der so-

genannten konvergenten und diskriminanten Validierung genutzt. Ausgangspunkt dieser Prüfstrategie sind Aussagen darüber, mit welchen Leistungen in anderen Tests die Leistungen in dem zu validierenden Test systematisch kovariieren sollen und mit welchen nicht. Diese Aussagen werden dann als Thesen empirisch überprüft. D.h. die Testergebnisse werden in den Kontext anderer Testergebnisse zu ähnlichen (konvergente Validität) und zu unähnlichen Konstrukten (diskriminante Validität) eingeordnet. Zur konvergenten Validierung des BIS-r-DGP Tests, einem Intelligenztest, wurde in der nachfolgenden Studie ein anderer Intelligenztest sowie ein Wissenstest herangezogen. Dabei wurde angenommen, dass der BIS-r-DGP Test substantiell mit diesen beiden Tests kovariiert. Demgegenüber sollten die Leistungen im BIS-r-DGP Test weitgehend unabhängig von Persönlichkeitsmerkmalen im engeren Sinne sowie von dem wahrgenommen Ausmaß an sozialer Unterstützung sein. Zur Erfassung dieser Konstrukte wurden Fragebogen eingesetzt.

4. Eine Studie zur konvergenten und diskriminanten Validität sowie zur Kriteriumsvalidität des BIS-r-DGP Tests

4.1 Untersuchungsteilnehmer

An der Studie nahmen insgesamt 379 Studierende teil, darunter 207 Frauen. Das Durchschnittsalter betrug 24,2 Jahre (Median 24), die Studierenden studierten im Mittelwert seit sieben Semestern. Die Gruppe setzte sich aus Studierenden unterschiedlicher Fächer zusammen, wobei die Studierenden der Wirtschaftswissenschaften überwogen. Allerdings haben nicht alle Teilnehmenden alle Instrumente bearbeitet bzw. Angaben zu den Kriterien getroffen, die exakten Teilnehmerzahlen sind in den nachfolgenden Abschnitten pro Test/pro Kriterium im Text oder in den Graphiken/Tabellen angegeben.

4.2 Eingesetzte Instrumente

Der zu validierende Test war der BIS-r-DGP Test (Kersting und Beauducel, 1997); dieser umfasst in seiner ursprünglichen Fassung 38 Aufgabentypen. In der vorliegenden Studie wurde auf den Einsatz der neun Aufgaben zur Merkfähigkeit verzichtet. Mit den verbleibenden Aufgaben wurden die operativen Fähigkeiten Verarbeitungskapazität, Einfallsmenge und Bearbeitungsgeschwindigkeit sowie die inhaltsgebundenen Fähigkeiten sprachgebundenes Denken, zahlengebundenes Denken und anschauungsgebundenes Denken erfasst (siehe Tabelle 1). Mittlerweile wird bei der DGP eine reduzierte Version des Tests eingesetzt (A2 Test), der sich auf die Erfassung der Verarbeitungskapazität mit verbalem und numerischem Aufgabenmaterial konzentriert. Die hier

präsentierten Ergebnisse zur Konstruktvalidierung können auf den aktuellen A2 Test übertragen werden, sofern es um die Dimension Verarbeitungskapazität sowie um den Umgang mit Sprache und Zahlen geht.

Tabelle 1: Beschreibung der sieben intellektuellen Fähigkeiten – Berliner Intelligenzstrukturmodell nach Jäger (1984)

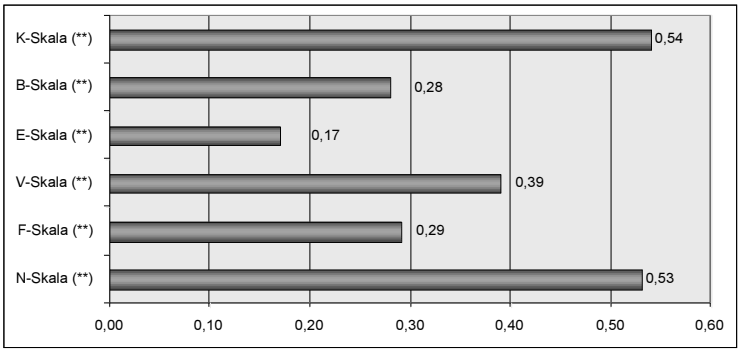
1. Operative Fähigkeiten		
K	**Verarbeitungskapazität**	Verarbeitung komplexer Informationen bei Aufgaben, die nicht auf Anhieb zu lösen sind, sondern Heranziehen, Verfügbarhalten, vielfältiges Beziehungsstiften, formallogisch exaktes Denken und sachgerechtes Beurteilen von Informationen erfordern.
E	**Einfallsreichtum**	Flüssige, flexible und auch originelle Ideenproduktion, die Verfügbarkeit vielfältiger Informationen, Reichtum an Vorstellungen und das Sehen vieler verschiedener Seiten, Varianten, Gründe und Möglichkeiten von Gegenständen und Problemen voraussetzt, wobei es um problemorientierte Lösungen geht, nicht um ein ungesteuertes Luxurieren der Phantasie.
M	**Merkfähigkeit**	Aktives Einprägen und kurz- oder mittelfristiges Wiedererkennen oder Reproduzieren von verbalem, numerischem oder figural-bildhaftem Material.
B	**Bearbeitungs-geschwindigkeit**	Arbeitstempo, Auffassungsleichtigkeit und Konzentrationskraft beim Lösen einfach strukturierter Aufgaben von geringem Schwierigkeitsniveau.
2. Inhaltsgebundene Fähigkeiten		
V	**Sprachgebundenes Denken**	Einheitsstiftendes Merkmal ist hier das Beziehungssystem Sprache. Ein dem Grad seiner Aneignung und Verfügbarkeit entsprechendes Fähigkeitsbündel scheint bei allen sprachgebundenen Operationen mitbestimmend zu sein.
N	**Zahlengebundenes Denken**	Analog zu V kann hier der Grad der Aneignung und Verfügbarkeit des Beziehungssystems Zahlen als einheitsstiftendes Merkmal interpretiert werden.
F	**Figural-bildhaftes Denken**	Diese Einheit ist zunächst durch die benannte Gemeinsamkeit des Aufgabenmaterials charakterisiert.

Zur konvergenten Validierung des BIS-r-DGP Tests wurde bei 320 Personen die deutschsprachige Fassung eines in den USA sehr verbreiteten Verfahrens zur Messung der Intelligenz eingesetzt (im Folgenden als „US-amerikanischer Intelligenztest“ bezeichnet). Der Einsatz eines international etablierten Tests erleichtert die Einordnung von BIS-r-DGP Ergebnissen in den Kontext internationaler Forschungsergebnisse. Darüber hinaus wurde bei 232 Personen der Wissenstest aus dem Intelligenz-Struktur-Test 2000 R (Amthauer, Brocke, Liepmann und Beauducel, 2001) verwendet.

4.3 Diskriminante und konvergente Validität

Die substantiellen Korrelationen (siehe Graphiken 1 und 2) zwischen der BIS-r-DGP Dimension Verarbeitungskapazität und den inhaltsgebundenen Fähigkeiten einerseits sowie dem US-amerikanischen Intelligenztest und dem IST Wissenstest andererseits unterstreichen die Konstruktvalidität des BIS-r-DGP Tests. Die Verarbeitungskapazität ist auch beim Lösen der Aufgaben eines anderen Intelligenztests gefragt. Da sowohl der IST-2000 R Wissenstest als auch der US-amerikanische Intelligenztest Aufgaben mit Zahlen, Figuren und Worten nutzen, kommt es darüber hinaus zu einer systematischen Kovariation mit den BIS-r-DGP Inhaltsklassen. Die Bearbeitungsgeschwindigkeit ist bei dem US-amerikanischen Intelligenztest ebenso gefordert, nicht aber so sehr bei dem weniger „gespeedeten“ Wissenstest. Der Einfallsmenge kommt plausiblerweise insgesamt eine geringere Bedeutung zu, insbesondere bei der Abfrage von Wissen spielt sie keine Rolle.

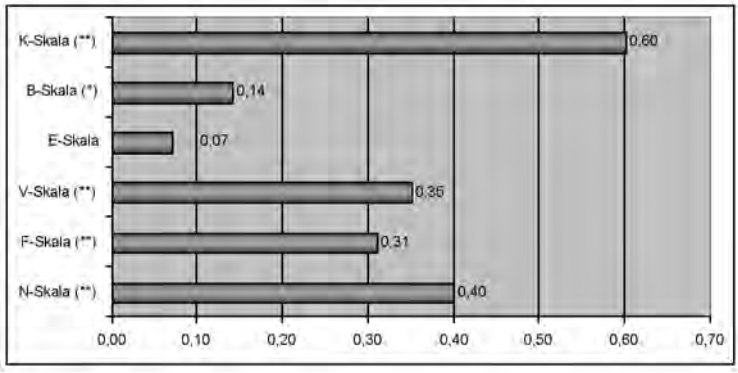
Graphik 1: Korrelationen zwischen dem BIS-r-DGP-Test und einem US-amerikanischen Intelligenztest (N=320)



Erläuterungen: ** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.
US-amerikanischen Intelligenztest: Cronbachs Alpha: .83

B-Skala = Bearbeitungsgeschwindigkeit; E-Skala = Einfallsmenge; K-Skala = Verarbeitungskapazität;
N-Skala = Zahlengebundenes Denken; V-Skala = Sprachgebundenes Denken; F-Skala = Figural-bildhaftes Denken

Graphik 2: Korrelationen zwischen dem BIS-r-DGP Test und IST-Wissenstest, Gesamtwert (N=232)

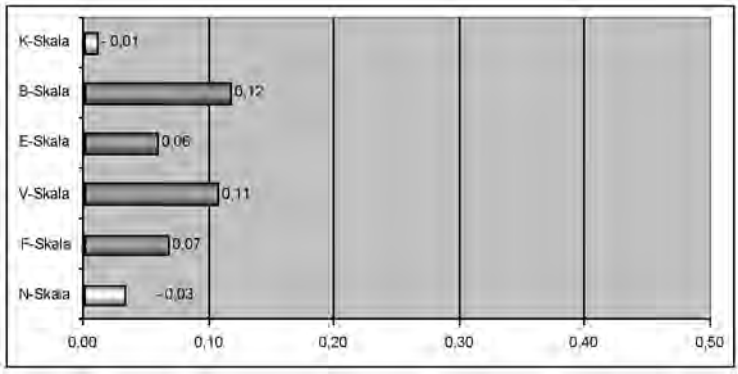


Erläuterungen: * Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.
** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.
IST 2000 R, Wissen Gesamtskala Cronbachs Alpha .93, Amthauer, R., Brocke, B., Liepmann, D., Beauducel, A. (2001)

B-Skala = Bearbeitungsgeschwindigkeit; E-Skala = Einfallsmenge; K-Skala = Verarbeitungskapazität;
N-Skala = Zahlengebundenes Denken; V-Skala = Sprachgebundenes Denken; F-Skala = Figural-bildhaftes Denken

Im Rahmen der diskriminanten Validierung zeigt sich – wie erwartet – dass das Ergebnis im BIS-r-DGP Intelligenztest nicht mit der wahrgenommenen sozialen Unterstützung im Zusammenhang steht (siehe Graphik Nr. 3, keine signifikante Korrelation).

Graphik 3: Korrelationen zwischen dem BIS-r-DGP Test und einer Skala zur wahrgenommenen Sozialen Unterstützung (N=253)



Erläuterungen: Keine Korrelation ist signifikant.
Fragebogen zur sozialen Unterstützung (FSozU-K, Sommer, Fydrich, Menzel, Höll, 1987)

B-Skala = Bearbeitungsgeschwindigkeit; E-Skala = Einfallsmenge; K-Skala = Verarbeitungskapazität;
N-Skala = Zahlengebundenes Denken; V-Skala = Sprachgebundenes Denken; F-Skala = Figural-bildhaftes Denken

Ebenfalls im Sinne der diskriminanten Konstruktvalidität zeigt sich, dass die mit dem „BIS-r-DGP“ Test erfassten Fähigkeiten nur gering mit den Persönlichkeitsmerkmalen im Sinne des Fünf-Faktoren-Modells korrelieren (siehe Tab. 2). Die nominell am deutlichsten ausgeprägten Werte in Höhe von .22 finden sich theoriekonform zwischen der Gewissenhaftigkeit und der Bearbeitungsgeschwindigkeit sowie zwischen der Extraversion und der Einfallsmenge. Ein leichter negativer Zusammenhang ergibt sich zwischen der Verarbeitungskapazität und dem Neurotizismus – anders formuliert: Emotional stabile Personen erzielen in der Testsituationen etwas bessere Werte.

Tabelle 2: Korrelationen zwischen dem BIS-r-DGP Test und einer Kurzskala zur Erfassung von fünf Persönlichkeitsfaktoren (N=317)

	Extra-version	Verträglich-keit	Gewissen-haftigkeit	Neuro-tizismus	Offenheit für Neues
K-Skala	- 0,14 (*)	- 0,07	- 0,06	- 0,20 (**)	- 0,05
B-Skala	0,06	- 0,02	0,22 (**)	- 0,09	- 0,03
E-Skala	0,22 (**)	- 0,13 (*)	0,05	- 0,08	0,11
V-Skala	0,08	- 0,10	0,12 (*)	- 0,19 (**)	0,10
F-Skala	0,08	- 0,07	0,03	- 0,14 (*)	0,02
N-Skala	- 0,15 (**)	- 0,03	0,03	- 0,11	- 0,18 (**)

Erläuterungen: * Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.
** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.
Kurzversion des Big Five Inventory (Rammstedt & John, 2005; 2007).

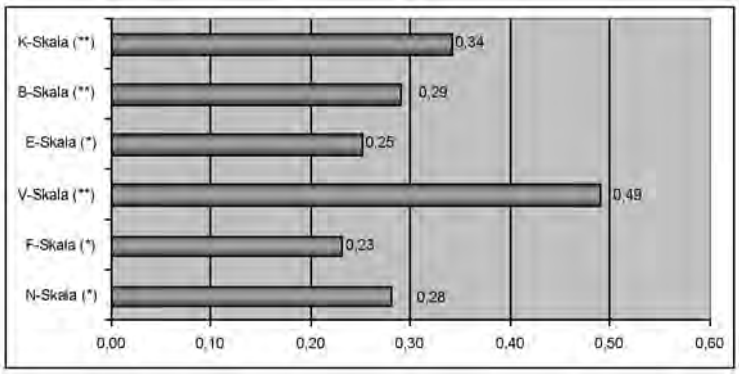
B-Skala = Bearbeitungsgeschwindigkeit; E-Skala = Einfallsmenge; K-Skala = Verarbeitungskapazität;
N-Skala = Zahlengebundenes Denken; V-Skala = Sprachgebundenes Denken; F-Skala = Figural-bildhaftes Denken

4.4 Kriteriumsvalidität

Im Rahmen der Studie sollte auch die Kriteriumsvalidität des Tests geprüft werden.

Als erstes Kriterium wurde der Zusammenhang mit der Abiturnote der Testteilnehmer(innen) bestimmt, die Ergebnisse sind in Graphik Nr. 4 dargestellt. Alle Skalen des BIS-r-DGP Tests korrelieren substantiell mit der Abitur-Note, der Zusammenhang zeigt sich insbesondere für die Inhaltsklasse „sprachgebundenes Denken“ sowie für die „Verarbeitungskapazität“.

Graphik 4: Korrelationen zwischen dem BIS-r-DGP Test und der Abitur-Note (N=85)

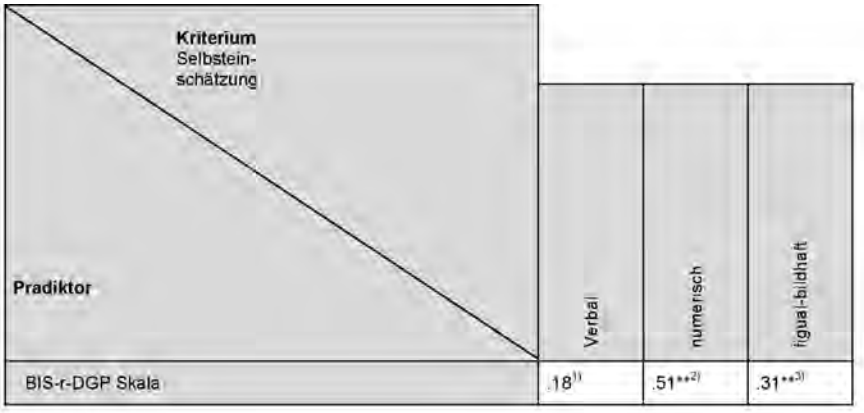


Erläuterungen: ** Die Korrelation ist auf dem Niveau von 0,01 (2-seitig) signifikant.
* Die Korrelation ist auf dem Niveau von 0,05 (2-seitig) signifikant.

B-Skala = Bearbeitungsgeschwindigkeit; E-Skala = Einfallsmenge; K-Skala = Verarbeitungskapazität; N-Skala = Zahlengebundenes Denken; V-Skala = Sprachgebundenes Denken; F-Skala = Figural-bildhaftes Denken

Darüber hinaus wurde anhand der Daten von 102 Personen der Zusammenhang zwischen den BIS-r-DGP Testleistungen und Selbsteinschätzungen bestimmt. Dazu lieferten die Testteilnehmer vor (!) der BIS-r-DGP Testung auf sechsfach abgestuften Likert-Single-Item-Skalen Selbsteinschätzungen zu ihren Fähigkeiten im Umgang mit Aufgaben mit (1) verbalem, (2) numerischem und (3) figuralem Material. Trotz aller Vorbehalte gegenüber Selbsteinschätzungen gilt, dass mit derartigen Maßen systematische Varianz erfasst wird, die in substantieller Beziehung zu relevanten Außenvariablen steht. So korrelieren Selbsteinschätzungen beispielsweise mit Vorgesetzteneinschätzungen. Die technische Qualität (Reliabilität) von Selbsteinschätzungen ist sehr zufrieden stellend (Harris und Schaubroek, 1988).

Tabelle 3: Kriteriumsvalidität: Korrelation der drei Inhaltsklassen des BIS-r-DGP Tests mit der vor der Testung erhobenen Selbsteinschätzung zu den mit diesen Modulen homologen Dimensionen (N = 102)



Erläuterungen: ** p≤.01.
Skalen (Prädiktoren):
1) = BIS-r-DGP sprachgebundenes Denken
2) = BIS-r-DGP zahlengebundenes Denken
3) = BIS-r-DGP figural-bildhaftes Denken

Mit Ausnahme des Zusammenhangs zwischen der Selbsteinschätzung und der tatsächlichen Testleistung in der BIS-r-DGP Skala für den Umgang mit verbalem Aufgabenmaterial (sprachgebundenes Denken) sind alle Zusammenhänge signifikant. Für den Bereich, der einer Selbsteinschätzung besonders gut zugänglich ist, nämlich den Umgang mit Zahlen, zeigt sich erwartungsgemäß ein besonders hoher Koeffizient. Der ausbleibende Zusammenhang zwischen der BIS-r-DGP Skala sprachgebundenes Denken und der Selbsteinschätzung der Fähigkeit, mit verbalem Aufgabenmaterial umzugehen, ist eventuell darauf zurückzuführen, dass bei der Beurteilung sprachlicher Fähigkeiten ein größerer Interpretationsspielraum besteht. Wer nicht gut rechnen kann, wird dies in seinem Leben immer wieder erfahren haben und kann sich entsprechend treffsicher einschätzen. Demgegenüber korrelieren Selbst- und Fremdeinschätzung bei sprachlichen Leistungen geringer.

5. Fazit

Die Studie zur Konstruktvalidität belegt gemeinsam mit den Ergebnissen früherer Studien (z. B. Kersting und Beauducel, 1997), dass mit dem BIS-r-DGP Test, bzw. mit den Aufgaben dieses Tests, die mittlerweile Bestandteil des A-2 Tests sind, die in den Bedeutungshorizont des wissenschaftlichen Begriffs “Intelligenz” fallenden Fähigkeiten erfasst werden, die mit dem Test/mit den Aufgaben erfasst werden sollen. Damit können auch zahlreiche Befunde zur Kriteriumsvalidität von Intelligenztests auf den BIS-r-DGP Test übertragen werden: Es kann davon ausgegangen werden, dass sich mit diesem Test/mit diesen Aufgaben der Erfolg bei kognitiv geprägten Lebensaufgaben wie Schule, Ausbildung, Studium und Beruf sehr gut vorhersagen lässt. Dies zeigt sich auch anhand der Korrelationen mit den Schulnoten im Abitur und den Selbsteinschätzungen. Die Einbettung der mit dem Test gemessenen Fähigkeiten ins Netz anderer Fähigkeiten und Persönlichkeitsmerkmale ermöglicht eine klare Interpretation der Testbefunde und ist Grundlage für die Planung von Eignungsuntersuchungen sowie für die Konstruktion neuer Verfahren und neuer Verfahrenskombinationen.
Wünschen wir der DGP, in unserem Interesse, dass sie sich in den nächsten Jahrzehnten weiterhin für die Testdiagnostik engagiert.

6. Literatur

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). Intelligenz-Struktur-Test 2000 R. Göttingen: Hogrefe.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), Intelligence. Measurement, theory, and public policy (pp. 147171). Urbana, IL: University of Illinois Press.

Cronbach, L. J., & Meehl, P. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

DIN (2002). DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen. Berlin: Beuth.

Harris, M. M. & Schaubroek, J. (1988). A meta-analysis of self supervisor, self peer and peer supervisor ratings: Personnel psychology, 41, 43-62.

Höfer, Th., Przyembel, H. & Verleger, S. (2004). New evidence for the theory of the stork. Paediatric and Perinatal Epidemiology, 18, 88-92.

Hülsheger, U. R., Maier, G. W. & Stumpp, T. (2007). Validity of general mental ability for the prediction of job performance and training success in Germany: A meta-analysis. International Journal of Selection and Assessment, 15, 3-18.

Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 7298.

Jäger, A.O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. Psychologische Rundschau, 35, 21-35.

Jäger, A.O. (1986). Validität von Intelligenztests. Diagnostica, 32, 272-289.

Kersting, M. & Beauducel, A. (1997). Der neue DGP-Leistungstest »BIS-r-DGP«: Informationen zu ausgewählten Testgütekriterien und zur Normierung. DGP Informationen, 46, Heft 55, 92-102.

Kramer, J. (2009). Allgemeine Intelligenz und beruflicher Erfolg in Deutschland. Vertiefende und weiterführende Metaanalysen. Psychologische Rundschau, 60, 82-98.

Messick S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 3345). Hillsdale, NJ: Erlbaum.

Rammstedt, B. & John, O. P. (2005). Kurzversion des Big Five Inventory (BFI-K). Entwicklung und Validierung eines ökonomischen Inventars zur Erfassung der fünf Faktoren der Persönlichkeit. Diagnostica, 51, 195-206.

Rammstedt, B. & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. Journal of Research in Personality, 41, 203-212.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C. & De Fruyt, F. (2003a). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. Personnel Psychology, 56, 573-605.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., De Fruyt, F. & Rolland, J. P. (2003b). A meta-analytic study of general mental ability validity for different occupations in the european community. Journal of Applied Psychology, 88, 1068-1081.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta analysis and cumulative knowledge in psychology. American Psychologist, 47, 1173-1181.

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, 262-274.

Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, H. R., Pearlman, K. & McDaniel, M. (1993). Refinements in validity generalization methods: Implication for the situational specificity hypothesis. Journal of Applied Psychology, 78, 3-12.

Sommer G, Fydrich T, Menzel U, Höll B (1987) Fragebogen zur sozialen Unterstützung (Kurzform: SOZU-K-22). Zeitschrift für klinische Psychologie 16, 434-436

Wiggins, J. S. (1973). Personality and prediction: Principles of personality assessment. Reading, MA: AddisonWesley.

Wittmann, W. W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J.R. Nesselroade & R.B. Cattell (Eds.), Handbook of multivariate experimental psychology. (2nd ed., pp. 505-560). New York: Plenum.

Korrespondenzanschrift des Autors:

Prof. Dr. Martin Kersting
Martin@kersting-internet.de

Für mehr Informationen siehe:
www.Kersting-internet.de