Kersting, M. (2007). Wenn Tests in die Jahre kommen. Probleme des Einsatzes überalteter Testverfahren.
In C. Lorei (Hrsg.), Polizei und Psychologie (S. 565-577). Frankfurt: Verlag für Polizeiwissenschaft.

in:
Clemens Lorei (Hrsg.)
2007

Polizei & Psychologie

Kongressband der Tagung „Polizei & Psychologie"
am 3. und 4. April 2006 in Frankfurt am Main

Schriftenreihe Polizei & Wissenschaft

ISSN 1610-7500
ISBN 978-3-935979-84-9

Verlag für Polizeiwissenschaft
Dr. Clemens Lorei

Wenn Tests in die Jahre kommen: Probleme des Einsatzes überalterter Testverfahren

Martin Kersting

1 Einleitung

Sowohl im Kontext der Personalauswahl als auch im Kontext von Beförderungsentscheidungen nutzt die Polizei in vielen Bundesländern standardisierte psychologische Testverfahren zur Messung kognitiver Kompetenzen. Dies ist sinnvoll, da solche Verfahren nachweislich treffsichere Ergebnisse liefern und alternativen Verfahren (z. B. Interviews und Assessment Centern) hinsichtlich Objektivität und Gültigkeit weit überlegen sind (siehe z. B. Schmidt und Hunter, 1998). Die Qualität der verfügbaren Testverfahren variiert allerdings erheblich. Im Fokus des vorliegenden Beitrags steht zunächst die Frage, inwieweit das Alter eines Testverfahrens sich negativ auf dessen Testqualität auswirkt. Diese Frage ist relevant, da einige Testverfahren über viele Jahre (teilweise Jahrzehnte) hinweg unverändert genutzt werden. Während das Thema im zweiten Abschnitt theoretisch erörtert wird, werden im dritten Abschnitt kurz die empirischen Ergebnisse skizziert, die am Beispiel der alten Version des Wilde-Intelligenz-Tests (WIT, Jäger & Althoff, 1994) die Wirkung der Überalterung von Testverfahren auf die Testqualität aufzeigen. Im vierten und letzten Abschnitt wird konkret dargelegt, wie die Erfahrungen mit einem Test bei dessen Überarbeitung und Modernisierung konstruktiv genutzt werden können.

2 Probleme des Einsatzes überalterter Testverfahren

In der Regel ist ein „neues" Auto leistungsstärker als ein „altes", vergleichbares gilt für viele technische Geräte vom Telefon bis zum Fernseher. Das Vertrauen, das Patienten ihrem Arzt entgegenbringen, sinkt, wenn die diagnostischen Geräte in der Praxis den Charme vergangener Jahrzehnte ausstrahlen. Wie aber verhält es sich mit standardisierten psychologischen Testverfahren? Bedeutet „alt" hier immer schlecht und „neu" immer gut? „Alter" per se ist bei Tests weder ein Qualitätsmangel, noch ein Gütesiegel.

Allerdings müssen „alte" Tests regelmäßig überprüft werden, um nachzuweisen, dass sie auch aktuell noch funktionstüchtig sind. Grundsätzlich ist es erstrebenswert, dass mit einem Test umfassende Erfahrungen gemacht und Daten gesammelt werden. Der Ernstfalleinsatz eines neuen Tests, zu dem keine Daten vorliegen, ist kaum zu rechtfertigen. Der Umfang der zur Verfügung stehenden Testdaten variiert in Abhängigkeit von der Testeinsatzhäufigkeit. Auch in kurzer Zeit kann ein Test bei sehr vielen Personen administriert werden, wie die PISA Studie zeigt (Deutsches PISA Konsortium, 2001). Zu den PISA Testaufgaben liegen belastbare Daten vor, obwohl der Test relativ jung ist. In der Regel kovariiert die zur Verfügung stehende Datenbasis aber mit dem Alter des Tests: Je älter ein Test ist, umso mehr Daten liegen zu seiner Bewährung vor. In dieser Hinsicht ist „alt" gleichbedeutend mit „erfahren" und im Falle, dass die Daten zufriedenstellend ausfallen, positiv zu werten. Die Veralterung eines Tests birgt aber verschiedene Gefahren mit sich. Im Folgenden soll auf Gefährdungen der Funktionstüchtigkeit und auf Einschränkungen der Interpretierbarkeit von Normwerten eingegangen werden. Ein weiterer kurzer Abschnitt skizziert die Regelungen der DIN 33430 zur Aktualität der Kennwerte von Testverfahren.

2.1 Gefährdung der Funktionstüchtigkeit

Besonders bei Testaufgaben, die explizit oder implizit Kenntnisse erfordern, besteht die Gefahr, dass das Antwortverhalten zeitspezifischen Besonderheiten unterworfen ist. Aktuell, nach den Irak-Kriegen, wissen mehr Menschen, dass Bagdad die Hauptstadt des Irak ist, als dies vor dem ersten Golfkrieg 1980 der Fall war. Davon sind sowohl Testitems betroffen, die unmittelbar Wissen abfragen (Wie heißt die Hauptstadt des Irak?) als auch Items, die ein solches Wissen indirekt voraussetzen. Beispielsweise könnte eine sprachliche Analogieaufgabe wie folgt aufgebaut sein:

Irak verhält sich zu Bagdad wie >?< zu England

Die Testteilnehmer sollen anstelle des Fragezeichens aus den Distraktoren
a) Paris b) Berlin c) London d) New York

eine Lösung auswählen. Für diese Aufgabe hätte man vor 1980 vermutlich andere Kennwerte erhalten als aktuell.

Auch Wortbedeutungen können sich über die Jahre hinweg zumindest in Nuancen verändern. Die Nutzungshäufigkeit und somit Vertrautheit von Wörtern ist nachweislich zeitlichen Schwankungen unterlegen. Dies wirkt sich auf Tests mit verbalem Aufgabenmaterial aus.

Neben unmittelbaren Kenntnissen sind auch Fertigkeiten wie Rechtschreibung, Rechnen und Fremdsprachen über die Zeit hinweg in den Bevölkerungskohorten unterschiedlich gut ausgebildet.

Zunächst schwer vorstellbar ist die Tatsache, dass sich auch die durchschnittliche Fähigkeit zum Umgang mit abstrakten Aufgaben über die Zeit hinweg verändert. Derartige Befunde werden unter dem Begriff des „Flynn-Effekts" diskutiert. James R. Flynn hatte gezeigt, dass jüngere Jahrgänge bei einem älteren und seitdem unverändert gebliebenen Test besser abschnitten als die früheren Generationen und zwar stiegen die durchschnittlichen Leistungen um ca. sieben IQ-Punkte pro Jahrzehnt. Den größten Zugewinn verzeichnete Flynn gerade bei den sprach- und >kulturfreien< Tests, die keinerlei Wissen verlangten. Eine mögliche Ursache hierfür könnte die unterschiedliche Vertrautheit mit dem abstrakten und häufig figuralen Aufgabenmaterial sein. Es zeichnete sich außerdem eine Korrelation zwischen dem Leistungsanstieg und der Tatsache ab, dass immer mehr Menschen immer länger die Schule besuchen.

Was immer die Ursachen sind: Die unterschiedliche Vertrautheit mit dem Testmaterial kann sich auf die Funktionsweise der Testaufgaben auswirken. So kann es sein, dass ein ursprünglich guter Test nach vielen Jahren oder Jahrzehnten nicht mehr differenziert, weil zu viele Personen alle Aufgaben richtig bearbeiten oder überhaupt keine Aufgabe lösen. Auch ist es möglich, dass ein Test, der einmal das logische Denken erfasst hat, über die Jahre zu einem Kenntnistest wird, weil das Aufgabenmaterial, mit dem die Testteilnehmer Schlussfolgerungen vollziehen sollen, nur noch wenigen Personen vertraut ist usw. Solche Qualitätseinbußen sind aber keinesfalls zwangsläufig, die Funktionsfähigkeit eines Tests muss im Einzelfall geprüft werden.

Schließlich steigt mit den Jahren der Testnutzung die Gefahr, dass der Testschutz nachlässt und die Testaufgaben bekannt werden, der Test „verbrennt", wie man umgangssprachlich sagt. Besonders problematisch ist es, wenn die Testaufgaben nur einigen Personen vorab bekannt sind, anderen aber nicht.

2.2 Einschränkungen der Interpretierbarkeit von Normwerten

Bei älteren Testverfahren muss den Bezugsgrößen, mit denen die individuell erreichten Ergebnisse verglichen werden, besondere Aufmerksamkeit gewidmet werden. Viele Tests bieten zur Interpretation der Testergebnisse an definierten Gruppen (Referenzgruppen) gewonnene Vergleichswerte (Normwerte) an. Bei der Interpretation der Normwerte wird berücksichtigt, wie stark ein Merkmal einer Person im Vergleich mit einer für die Eignungsaussage relevanten Stichprobe anderer Personen ausgeprägt ist. Diese Interpretation kann problematisch werden, wenn die aktuell getesteten Personen z. B. andere Bildungserfahrungen innerhalb und außerhalb der Schule gemacht haben als die Personen der Referenzgruppe.

So hat sich beispielsweise die durchschnittliche Ausprägung der Rechtschreibleistungen in der Bevölkerung innerhalb eines vergleichsweise kurzen Zeitabschnitts verändert – nämlich verschlechtert (Kersting und Althoff, 2004; Kiepe 1998). Die Anwendung älterer Normen würde unter diesen Umständen eine zu „strenge" Bewertung implizieren.

Bei anderen Tests, wie z. B. Tests zum schlussfolgernden Denken mit abstraktem Aufgabenmaterial, steigt, wie bereits oben erwähnt, die mit den Tests gemessene Intelligenzleistung über die Jahre hinweg an („Flynn-Effekt", siehe z. B. Horn und Bullheller, 2004; Rodgers, 1998). Während eine Überalterung von Normen für Rechtschreibtests somit mit großer Wahrscheinlichkeit zu „zu strengen" Normen führt, sind veraltete Normen für Intelligenztests vermutlich „zu mild".

2.3 Was sagt die DIN 33430 zur Veralterung von Tests?

In den vorherigen Abschnitten wurde aufgezeigt, dass ein Test mit den Jahren seine Funktionstüchtigkeit einbüßen und die Interpretierbarkeit der Normen gefährdet sein kann. Solche Qualitätseinbußen sind aber keinesfalls zwangsläufig. Möglich ist auch, dass ein Test und seine Normen über Jahrzehnte hinweg ihre Qualität bewahren. So wie ein Auto regelmäßig zum TÜV muss, muss auch ein Test regelmäßig auf seine Funktionstüchtigkeit überprüft werden. Die Qualitätsstandards für „Verfahren und deren Einsatz

bei berufsbezogenen Eignungsbeurteilungen" formuliert die DIN 33430 (DIN, 2002), die im Jahre 2002 verabschiedet wurde (siehe z. B. Kersting und Püttner, 2006). Gerade für die öffentliche Verwaltung und die Polizei kommt der DIN 33430 eine besondere Bedeutung zu, da sie nach Wegener (2003) faktisch – wenn auch nicht im engeren Sinn rechtlich – die Bedeutung so genannter normkonkretisierender Verwaltungsvorschriften hat. Durch diesen „Richtliniencharakter" ist sie mit „echten" Verwaltungsvorschriften vergleichbar. (Zur Frage der rechtlichen Verbindlichkeit der DIN Norm siehe auch Abeln und Reimann, 2004, sowie Kersting und Püttner, 2006). Die DIN 33430 schreibt vor, dass die Zuverlässigkeit und Gültigkeit eines Verfahrens sowie die Aktualität der Normen mindestens alle acht Jahre überprüft werden. Vorgeschrieben wird also lediglich eine Überprüfung der Kennwerte und Normen, keinesfalls muss ein Test zwangsläufig alle acht Jahre neu normiert werden, wie einige die DIN fehlinterpretieren. Ob die Verfahrenshinweise (Testmanuale) eines Tests den Anforderungen der DIN 33430 entsprechen, kann rasch und einfach durch die Anwendung der „Checkliste 1", der so genannten „DIN Screen" Publikation (Kersting, 2006) überprüft werden. Diese Checkliste, die auch online verfügbar ist (http://www.kersting-internet.de/DIN-Screen.html), gilt offiziell als „Standard zur Information und Dokumentation von Instrumenten zur Erfassung menschlichen Erlebens und Verhaltens des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen". Die Checklisten-Kontrolle der Verfahrenshinweise eines Tests erlaubt ein rasches Negativ-Screening eines Tests. Eine berufsbezogene Eignungsbeurteilung, bei der ein Verfahren eingesetzt wird, für das keine DIN-33430-kompatiblen Verfahrenshinweise vorliegen, ist in jedem Fall eine qualitativ unzureichende Eignungsbeurteilung nach DIN 33430.

3 Anwendungsbeispiel: Probleme des „alten" Wilde Tests

Welche Probleme beim Einsatz „alter" Testverfahren möglicherweise auftreten, soll nun konkret am Beispiel der alten Version des Wilde-Intelligenz-Tests (WIT, Jäger und Althoff, 1994) gezeigt werden. Der WIT gehört zu den etabliertesten Intelligenztestverfahren und wurde bzw. wird in zahlreichen Bundesländern von der Polizei genutzt. Entwickelt wurde er zwischen 1954 und 1962 von Jäger und Althoff in der Deutschen Gesellschaft für

Personalwesen (DGP). Nachdem der WIT zunächst fast zwanzig Jahre lang bei ca. 250.000 Personen bei der DGP im Rahmen von Eignungsuntersuchungen zum Einsatz kam, wurde er 1983 veröffentlicht und erschien 1994 in der zweiten, vermeintlich „revidierten" Auflage. In der revidierten Auflage waren aber weder aktualisierte Normen zu finden, noch konnte der Test mit revidiertem Aufgabenmaterial und ergänzenden Studien aufwarten. Gerade einmal vier der insgesamt 560 Items wurden in nur einem Subtest (ER) geringfügig verändert.

Knebelau (2002) hat anhand aktueller Daten von 382 Bewerber(inne)n um einen Ausbildungsplatz im gehobenen Dienst der öffentlichen Verwaltung geprüft, inwieweit sich der „alte" WIT aktuell noch bewährt. Dabei zeigte sich unter anderem Folgendes:

- Die psychometrischen Kennwerte einiger Aufgaben des alten WIT sind unbefriedigend. So erwiesen sich die Tests Analogien, Zahlenreihen und Buchstabenreihen insgesamt als zu leicht. Beispielsweise finden im Test Zahlenreihen durchschnittlich 83 % aller Personen, die ein Item in Angriff nehmen, auch die richtige Lösung. Die Trennschärfen der Items waren durchgängig niedriger als die im Handbuch berichteten Werte.

- Die als parallel bezeichneten Testformen des alten WIT unterscheiden sich bei einigen Aufgaben hinsichtlich der mittleren Itemschwierigkeit sowie hinsichtlich der Reihung der Items nach ansteigender Schwierigkeit deutlich voneinander. Dabei zeigt sich, dass die Items der einen Testform nicht nur seltener gelöst, sondern auch seltener bearbeitet werden als die Items der anderen Testform.

- Eine Interpretation der Struktur der mit dem alten WIT erhobenen Testdaten auf der Basis des theoretischen Testkonzepts, dem Thurstone Modell, gelingt nicht.

- Während die bisherigen Daten den „alten" WIT im kritischen Licht darstellen, zeigte sich bei der Überprüfung der aktuellen Gültigkeit und Treffsicherheit des alten Verfahrens eine positive Überraschung. Der gegen Streuungseinschränkungen des Prädiktors und Reliabilitätseinschränkungen des Kriteriums korrigierte Zusammenhang zwischen dem alten WIT Gesamtwert und der durchschnittlichen Zwischenprüfungsnote im Rahmen einer Ausbildung zum Verwaltungsinspektor (N=122) ergab eine logisch positive Korrelation, d. h., aufgrund des Testergebnisses konnte der Ausbildungserfolg treffsicher vorhergesagt werden.

Als Fazit der Untersuchung kann festgehalten werden, dass der alte WIT revisionsbedürftig, aber auch revisionswürdig war.

4 Erfahrungen mit „alten" Tests können für die Modernisierung des Tests genutzt werden

Die Analyse der mit alten Tests aktuell gewonnenen Daten kann, wie im Falle des alten WIT, Schwächen des Tests aufdecken. Diese Ergebnisse können dann genutzt werden, um einen Test zielgerichtet zu modernisieren, wobei die Stärken des Tests gewahrt und ausgebaut werden, während den Schwächen des Tests Abhilfe geleistet wird. Exakt dies ist mit dem alten WIT geschehen. Aufgrund umfangreicher Analysen zum alten WIT sowie aufgrund neuer theoretischer und psychometrischer Entwicklungen wurde der neue, vollständig überarbeitete WIT-2 (Kersting, Althoff und Jäger, in Druck) entwickelt und neu normiert.

Zunächst wurde geprüft, welche der 14 Aufgaben des alten WIT revisionswürdig waren. Die besten sieben der 14 Aufgaben wurden ausgewählt, um nach einer grundsätzlichen Überarbeitung auch im neuen WIT enthalten zu sein. Die sieben Aufgaben verteilen sich auf zwei Gruppen. Bei den Aufgaben der Gruppe 1 handelte es sich um Aufgaben mit figuralem und numerischem Aufgabenmaterial. Diese „alten" Aufgaben funktionierten auch aktuell teilweise noch sehr gut. Da der alte WIT über zwei Testformen verfügte, standen pro Aufgabe doppelt so viele Items zur Verfügung, als für die Erstellung einer neuen Form notwendig waren. So konnte eine Bestenauswahl der Items erfolgen. Bei den Aufgaben der Gruppe 2 handelte es sich um Aufgaben mit sprachlichem Aufgabenmaterial, die einer gründlichen Überarbeitung bedurften.

Zusätzlich zu den sieben Aufgaben aus dem alten WIT wurden vier Aufgaben komplett neu entwickelt. Der neue WIT-2 umfasst somit sieben

Aufgaben, die nach gründlicher Überarbeitung und Modernisierung aus dem alten WIT übernommen wurden und vier komplett neuentwickelte Aufgaben. Insgesamt liegen 11 Aufgaben mit 243 Items vor (siehe Tabelle 1). Der Test ist modular aufgebaut, je nach Anforderungsprofil können gezielt einzelne Module allein oder in Kombination miteinander genutzt werden. Die Testzeit variiert entsprechend des Umfangs der genutzten Module zwischen 26 und 147 Minuten (siehe Abbildung 1).

Tabelle 1: Dimensionen und Testaufgaben des neuen Wilde-Intelligenz-Tests (WIT-2)

Dimension *(jede Dimension kann separat erfasst werden)*	***Testaufgaben***	***Itemzahl***	***Zeitbedarf*** *(Instruktion u. Laufzeit)*
Sprachliches Denken	(1) Analogien[2], (2) Gleiche Wortbedeutungen[2]	40	12 Min.
Rechnerisches Denken	(1) Grundrechnen[1], (2) Eingekleidete Rechenaufg.[2]	40	27 Min.
Räumliches Denken	(1) Abwicklungen[1], (2) Spiegelbilder[1]	40	22 Min.
Schlussfolgerndes Denken	(1) Analogien[2], (2) Abwicklungen[1], (3) Zahlenreihen[1]	60	14 Min. oder 35 Min. [4]
Merkfähigkeit	Merkfähigkeit[3]	21	9 Min.[5]
Arbeitseffizienz	Emails bearbeiten[3]	42	19 Min.
Wissen Wirtschaft	Wissen Wirtschaft[3]	20	5 Min.
Wissen Informationstechnologie	Wissen Informationstechnologie[3]	20	5 Min.
Die allgemeine Instruktion / Testeinführung dauert ca.			14 Min.
Nach ca. 90 Minuten Testung erfolgt eine Pause im Umfang von ca.			20 Min.
Gesamttestzeit (falls alle Module eingesetzt werden sollen)			ca. 147 Min.

[1] gegenüber dem "alten" WIT kaum verändert; [2] gegenüber dem "alten" WIT deutlich modifiziert; [3] vollständige Neuentwicklung; [4] als separates Modul: 35 Min., als Ergänzung (nur Zahlenreihen) zu den zu den sprachlichen, rechnerischen und figuralen Aufgaben: 14 Min. zusätzlich; [5] zwischen Einprägen und Wiedergabe wird eine andere Testaufgabe im Umfang von ca. 18 Minuten bearbeitet

Das Ziel der Revision und Neukonstruktion war die Entwicklung eines modernen, modular aufgebauten, psychometrisch hochwertigen und zugleich anwendungsorientierten Tests. Der WIT-2 zielt vorrangig auf die berufsbezogene Diagnostik, die Testaufgaben sind teilweise unmittelbar in eine Semantik aus dem Berufs- und Arbeitsleben eingekleidet. Der Test fokussiert nicht so sehr das abstrakte Denkvermögen, sondern berufliche Schlüsselqualifikationen und Grundfertigkeiten. Für jedes Modul stehen aktuelle, bildungs- und altersdifferenzierte Normdaten von mindestens 2.000 Personen zur Verfügung. Einzelne Aufgaben des neuen WIT-2 wurden insgesamt bei über 42.000 Personen eingesetzt. Der Berufsorientierung entsprechend, zielen die Normdaten nicht auf Bevölkerungsrepräsentativität, sondern auf Repräsentativität für eignungsdiagnostisch relevante Gruppen (Bewerber, Rehabilitanden).

Es werden (1) drei bildungsspezifische und (2) sechs altersspezifische Normgruppen angeboten: (1a) Personen mit Abitur (1b) Personen ohne Abitur (1c) Gesamtgruppe (40% Abitur, 60% kein Abitur). (2a) 14-17 Jahre; (2b) 18 Jahre; (2c) 19-22 Jahre; (2d) 23-27 Jahre; (2e) 28 und älter sowie (2f) Gesamtgruppe (ohne Altersdifferenzierung). Somit stehen insgesamt 3 * 6 = 18 Normen pro Aufgabe zur Verfügung. Bei der Erhebung der Daten stand der Praxisbezug im Vordergrund. Üblicherweise werden in die Normierung von Tests Daten einbezogen, die anhand von Forschungsuntersuchungen gewonnen werden. Die Teilnehmer bearbeiten die Tests häufig anonymisiert. Der überwiegende Teil der Normdaten für den WIT-2 wurde demgegenüber im Kontext des Ernstfalls von beruflichen Bewerbungssituationen erhoben.

Als theoretischer Ausgangspunkt der Konstruktion des „alten" Wilde Tests galt Thurstones Primary Mental Ability-Modell. Der neue WIT-2 repräsentiert fünf der sieben Primärfähigkeiten Thurstones (verbal comprehension, number, space, reasoning und memory). Darüber hinaus werden die Dimensionen Arbeitseffizienz und Wissen (Wirtschaft sowie Informationstechnologie) erfasst. Die statistischen Analysetechniken und die Intelligenzmodelle haben sich seit Thurstone weiterentwickelt. Den aktuellen Entwicklungen der Intelligenzstrukturforschung folgend, betrachtet der WIT-2 das Primärfaktorenmodell nicht als konträr zu hierarchischen Generalfaktorenmodellen, sondern sieht in diesen beiden Konzepten einander ergänzende Perspektiven, die in einem Modell mehrerer Generalitätsebenen überlappender

Faktoren integriert werden können. Das schlussfolgernde Denken wird im WIT-2 (anders als bei Thurstone) als eine dem verbalen, rechnerischen und räumlichen Denken übergeordnete Skala konzipiert.

Grundlegend neu ist die Aufgabe zur Erfassung der Arbeitseffizienz. Durch die Simulation einer Büroroutinetätigkeit (E-Mails bearbeiten) werden typische Anforderungen von Büroberufen (Ordnen, Sortieren, Vergleichen und Kontrollieren) in Form einer Arbeitsprobe nachgestellt. Anders als bei Aufmerksamkeitstests, geht es nicht um das Wahrnehmen von und Reagieren auf bereits bekannte, überlernte Reize, sondern um das effiziente Verarbeiten neuartiger Informationen. Die Testteilnehmer müssen, vergleichbar mit Mitarbeitern im Büro, einfache Regeln lernen und effizient anwenden.

Neu sind auch zwei Kenntnistests zu den beiden berufsrelevanten Wissensdomänen, Wirtschaft und Informationstechnologie.

Der WIT-2 erfasst die Merkfähigkeit und nicht, wie andere Tests, die unmittelbare Behaltensleistung. Die zu merkenden Informationen werden daher nicht unmittelbar nach dem Einprägen abgefordert, sondern nach der Einprägphase wird zunächst eine andere Aufgabe im Umfang von 18 Minuten bearbeitet.

Eine Besonderheit stellt schließlich die hohe Anwenderorientierung des Tests dar. Die fehlerfreie und komfortable Instruktion wird durch ein separates Instruktionsheft gewährleistet. Die Testauswertung sieht explizit anforderungsanalytisch gewichtete Testscores vor, um in jedem individuellen Anwendungsfall eine maßgeschneiderte Diagnostik zu gewährleisten. Der WIT-2 profitiert von über 50 Jahren Praxiserfahrung der Deutschen Gesellschaft für Personalwesen, die den Test herausgibt. Anspruch des WIT-2 ist es, ein Test von Praktikern für Praktiker sein.

Der WIT-2 hat sich nach Ausweis der psychometrischen Kennwerte sehr gut bewährt. Die internen Konsistenzen der Skalen liegen zwischen .78 und .95. Besonders wichtig sind die Informationen zur Gültigkeit des neuen Tests. Als Kriteriumsvalidität wird der Grad der Genauigkeit bezeichnet, mit dem von den Ergebnissen eines Tests auf ein Kriteriumsverhalten (z. B. den Ausbildungserfolg) geschlossen werden kann. Tests zur Personalauswahl

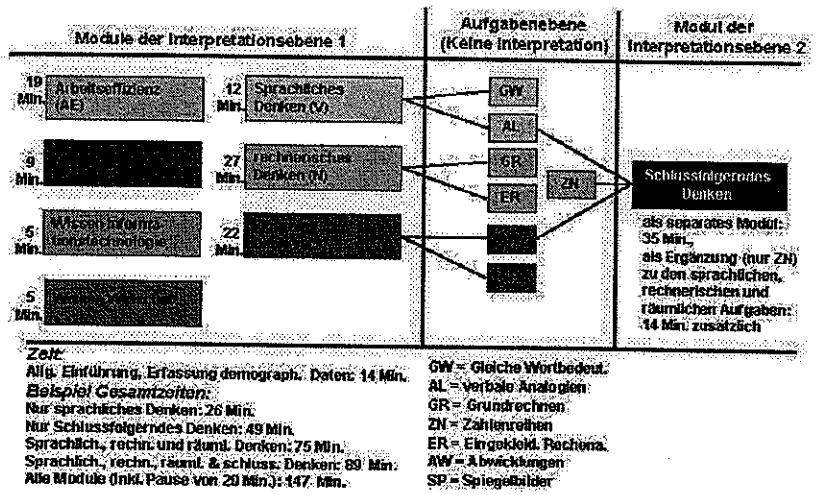

Abbildung 1: Überblick über die Module des neuen WIT-2

müssen in einem nachweisbaren Zusammenhang mit dem Ausbildungs- oder dem Berufserfolg stehen. Dieser Zusammenhang begründet die Vorhersagekraft der Testergebnisse. Zu jeder Aufgabe, die aus dem „alten" WIT in modifizierter Form übernommen wurde, wurden die bisherigen Ergebnisse zur Kriteriumsvalidität metaanalytisch zusammengefasst (pro Aufgabe 12 bis 20 unabhängige Studien mit einem N von 987 bis 2.277). Unter diesen Studien befanden sich allein vier Studien zur Bewährung des Tests bei der Polizei. Zur Prüfung der Gültigkeit des aktuellen Verfahrens wurden u.a. berufsnahe Kriterien, nämlich Noten in berufsorientierten Lehrgängen, herangezogen. Der WIT-2 erzielt hier gute prognostische Kriteriumsvaliditäten.

Bei der Konstruktvalidität geht es darum, inwieweit man von den Ergebnissen eines Tests auf den Ausprägungsgrad eines Konstrukts (z. B. einer Fähigkeit) schließen kann. Zur Überprüfung der Faktorenstruktur wurden u.a. konfirmatorische Faktorenanalysen gerechnet, dabei wurden gute Modellfits erzielt. Die Geltung der Konstruktannahmen ist nicht auf die Aufgaben

des neuen WIT-2 beschränkt, sondern zeigt sich auch in Untersuchungen mit Aufgaben (aus dem IST 2000 R und aus dem BIS-4 Test), die nicht zur Modellkonstruktion eingesetzt wurden. Insgesamt wurden die Beziehungen des neuen WIT-2 Tests zu 17 anderen Tests empirisch ausgelotet. Die Bereitstellung von teilnehmerorientierten Vorabinformationen zum Test sowie die motivational ansprechende Gestaltung der Testmaterialien dürfte dem Test eine sehr gute Akzeptanz bei den Testteilnehmer(inne)n sichern.

5 Zusammenfassung / Fazit

Wenn Tests in die Jahre kommen, ist das nicht grundsätzlich schlecht, aber es erfordert eine Überprüfung, ob der Test aktuell noch funktioniert. Der Beitrag skizziert auf theoretischer Ebene sowie anhand des Beispiels des alten WIT die möglichen Probleme veralteter Tests. Gleichzeitig zeigt das Beispiel des WIT wie die Analysen zu einem alten Test konstruktiv für die Überarbeitung und Neuentwicklung eines Tests genutzt werden können. Der neue WIT-2 profitiert von den umfassenden Erfahrungen, die mit dem alten WIT gemacht wurden und stellt zugleich ein modernes Diagnoseinstrument für die berufliche Eignungsdiagnostik dar.

6 Literatur

Abeln, C. & Reimann, G. (2004). DIN 33430 und die Folgen. Personalauswahl und -entwicklung im Umbruch. Arbeit und Arbeitsrecht, 11, 8-15.

Deutsches PISA Konsortium (Hrsg.) (2001) *PISA 2000. Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich.* Opladen: Leske und Budrich.

Horn, R. & Bulheller, St. (2004). Welche Folgen hat die Anwendung veralteter Normen? In: L. F. Hornke & U. Winterfeld (Hrsg.) *Eignungsbeurteilung auf dem Prüfstand* (S. 237-248). DIN 33430. Heidelberg: Spektrum Akademischer Verlag.

Jäger, A. O. & Althoff, K. (1994). *Der Wilde-Intelligenz-Test (WIT)* (Handanweisung, 2. revidierte Auflage). Göttingen: Hogrefe.

Kersting, M., Althoff, K. & Jäger, A. O. (in Vorbereitung). *Der Wilde-Intelligenz-Test 2 (WIT-2)* (vollständig überarbeitete, neunormierte Auflage). Göttingen: Hogrefe.

Kersting, M. (2006). *DIN Screen Leitfaden zur Kontrolle und Optimierung der Qualität von Verfahren und deren Einsatz bei beruflichen Eignungsbeurteilungen.* Lengerich: Pabst Science Publishers.

Kersting, M. & Althoff, K. (2004). *Rechtschreibungstest (RT).* Göttingen: Hogrefe.

Kersting, M. & Püttner, I. (2006). Personalauswahl: Qualitätsstandards und rechtliche Aspekte. In: H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie,* 2te Auflage (S. 841-861). Göttingen: Hogrefe.

Kiepe, K. (1998) Sieben Statements zur Ausbildungsreife. In W. Postal, K. Parmentier & K. Schober (Hrsg.), *Mangelnde Schulleistungen oder überzogene Anforderungen? Zur Problematik unbesetzt / unbesetzbarer Ausbildungsplätze* (S. 24 -30). Nürnberg: Bundesanstalt für Arbeit.

Knebelau, M. (2002). *Evaluation des Wilde-Intelligenz-Tests (WIT) und seiner Funktion als eignungsdiagnostisches Instrument.* Unveröffentlichte Diplomarbeit. Aachen: RWTH Aachen.

Rodgers, B. (1998). A critique of the Flynn Effect. Massive IQ gains, methodological artifacts, or both? *Intelligence, 26,* 337-356.

Schmidt, F.L. & Hunter, J.E. (1998). Messbare Personmerkmale: Stabilität, Variabili-tät und Validität zur Vorhersage zukünftiger Berufsleistung und berufsbezogenen Lernens. In: M. Kleinmann & B. Strauß (Hrsg.). *Potentialfeststellung und Personalentwicklung* (S. 16-43). Göttingen: Verlag für Angewandte Psychologie.

Wegener, M. (2003). Rechtliche Verbindlichkeit der DIN 33430 für Behörden und Gerichte. *DGP Informationen, 48,* Heft 57, S. 7-11.