Prof. Dr. Rolf R. Engel
Klinikum der Universität München, Klinik und Poliklinik für Psychiatrie und Psychotherapie
Nussbaumstr. 7, D-80336 München, ml-set@med.uni-muenchen.de

Hintergrundinformationen zur Visualisierung von Leistungstestergebnissen in der Internet-Applikation TDB2Online

Rolf R. Engel

Stand: 16.8.2018

Arten und Eigenschaften von Testwerten

Rohwerte sind die elementaren Werte, mit denen eine Leistung in einem Testverfahren beurteilt wird. Sie hängen von der Art der Aufgabe ab. Beispiele dafür sind:

- Anzahl richtiger Lösungen in einem Wissenstest (AW)
- Anzahl der Fehler in einer Kategorisierungsaufgabe (HCT)
- Zeit für die Durchführung in Sekunden bei einer visokonstruktiven Aufgabe (TMT)
- Anzahl richtig reproduzierter Wörter in einem Gedächtnistest (WL)
- Differenz zwischen der Anzahl richtig wiedererkannter Wörter minus Anzahl der fälschlich "wiedererkannten" Wörter in einer Wiedererkennensaufgabe (WL)
- Punkte in einer komplexen Aufgabe, die nach Richtigkeit und Schnelligkeit der Durchführung bewertet wird (MT)

Es ist unmittelbar zu sehen, dass sich mit Rohwerten Testleistungen zwischen verschiedenen Tests nicht vergleichen lassen. Die Punkte im Mosaiktest haben numerisch nichts mit den Zeiten im Trail-Making-Test zu tun, außerdem sind sie auch noch gegensätzlich gepolt.

Zum Vergleich zwischen Personen eignen sich Rohwerte hingegen durchaus, auch beim Vergleich von Gruppen (insbesondere, wenn sie hinsichtlich anderer Parameter vergleichbar sind) werden gerne Rohwerte verwendet.

Leistungswerte dienen dazu, die Leistungen über die verschiedenen Testverfahren und Rohwertarten hinweg vergleichbar zu machen. Sie liefern einen absoluten Maßstab zur Leistungsbeurteilung in einer standardisierten Form. Als erster hat David Wechsler dieses Prinzip angewendet, um einen Leistungsvergleich innerhalb der Subtests seiner Intelligenzbatterien zu ermöglichen (etwas, was mit den Rohwerten nicht möglich ist). Die Daten dafür kommen aus Normierungsuntersuchungen an Zufallsstichproben, bei denen einige hundert Personen pro Altersgruppe getestet werden. Die Standardisierung der Leistungswerte erfolgt an einer Gruppe junger Erwachsener etwa im Bereich zwischen 20 und 30 Jahren. Die Begründung für die Wahl dieser Gruppe liegt darin, dass in diesem Alter im Allgemeinen der individuelle Höhepunkt der kognitiven Leistungsfähigkeit in standardisierten Tests erreicht wird (ausgenommen sind Wissenstests, bei denen das Maximum später liegt). Bei jüngeren und bei älteren fällt die Leistung ab, allerdings bei unterschiedlichen kognitiven Leistungen nicht in gleicher Weise. Deshalb eignen sich andere Altersgruppen nicht zur Standardisierung, wenn man einen Vergleich von Leistungen über verschiedene Fähigkeiten hinweg haben will.

A**ltersnormierte Werte** wurden eingeführt, um die Interpretation der Leistungsfähigkeit eines einzelnen Patienten im Vergleich zu seiner Alterskohorte zu erleichtern. Auch bei ihnen handelt es sich um Standardwerte, genau wie bei den Leistungswerten. Der Unterschied liegt darin, dass die Bezugsbasis für die Standardisierung aus dem Teil der Normstichprobe

kommt, der altersmäßig mit dem Patienten vergleichbar ist. Diese Werte sind für eine absolute Leistungsbeurteilung über verschiedene Testverfahren hinweg kaum brauchbar, weil verschiedene Fähigkeiten sich im Altersverlauf unterschiedlich verändern. Sie sind aber eine wichtige Interpretationshilfe bei der Beurteilung der Leistung eines einzelnen Patienten im Vergleich zu seiner Alterskohorte. Die klinische Beurteilung im verbalen Testbefund stützt sich im Wesentlichen auf die altersnormierten Standardwerte. Die herkömmliche Auswertung vieler Testverfahren von Hand liefert neben den Rohwerten meistens nur altersnormierte Standardwerte, keine altersunabhängig standardisierten Leistungswerte. Man sollte sich trotzdem bewusst sein, dass mit altersnormierten Werten streng genommen die Ebene des Messens verlassen und die Ebene der Interpretation betreten wird.

Eigenschaften der benutzten Skala

Als Maßstab werden sowohl für die Leistungswerte als auch für die altersnormierten Standardwerte Skalen verwendet, die aus der Normalverteilung abgeleitet sind. Die Abszissenwerte der Normalverteilung selbst (*z-Werte*) eignen sich für die Kommunikation nicht gut (Kommawerte, negative Zahlen), weshalb seit Jahrzehnten nur daraus abgeleitete Maßstäbe benutzt werden. TDB2Online benutzt immer die IQ-Skalierung, bei der 100 der Mittelwert ist und 15 die Standardabweichung, sowohl für die Leistungswerte als auch für die altersnormierten Standardwerte.

Die Eigenschaften der IQ-Skala lassen sich an Hand der **Abbildung 1** erkennen. Am wichtigsten ist für die Interpretation der Vergleich zwischen den Werten der IQ-Skala und den Werten der Prozentrangskala: Man sieht, dass ein IQ-Wert von 130 einem Prozentrang von 97,7 entspricht, was nichts anderes bedeutet, als dass 97,7 Prozent der Vergleichsstichprobe einen niedrigeren IQ als 130 haben und die restlichen 2,3 Prozent einen höheren. Zu jedem IQ-Wert gibt es einen entsprechenden Prozentrangwert, den man ausführlicheren Tabellen entnehmen kann.

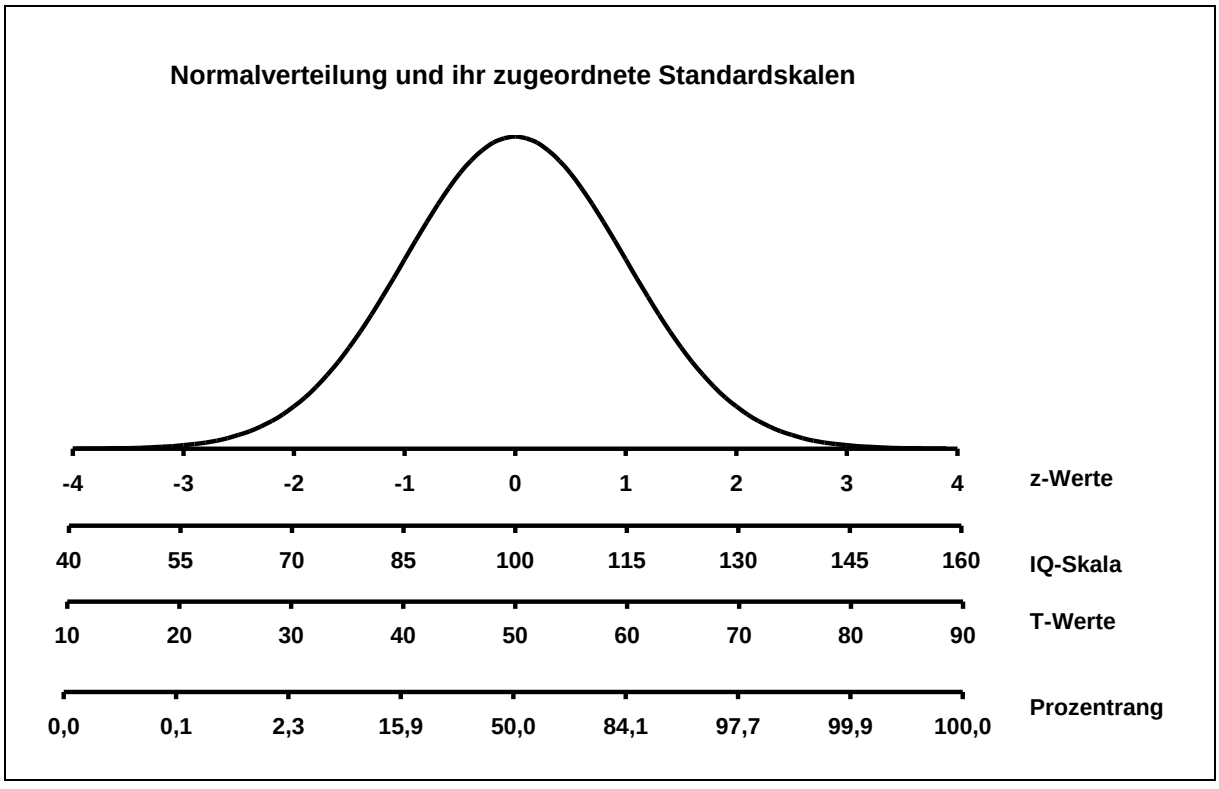


Abbildung 1: Normalverteilung und einige der von ihr abgeleiteten Standardskalen

Die IQ-Skala hat die angenehme Eigenschaft, dass sich leicht zu merkende und interpretativ sinnvolle verbale Beschreibungen für bestimmte Skalenabschnitte anbieten. Zwischen den IQ-Werten 90 und 110 liegen rund 50 Prozent der Normstichprobe. Dieser Bereich wird in den Befunden als "durchschnittlich" bezeichnet, gelegentlich auch mit zusätzlichen qualifizierenden Bemerkungen ("unterer Durchschnittsbereich", "am oberen Rand des Durchschnitts", o. ä.). Der IQ-Bereich von 80-90 umfasst rund 16 Prozent und wird als "niedrig" o.ä. bezeichnet, darunter liegt mit 70-80 der "sehr niedrige" Bereich (rund 7 %), und noch darunter (unter 70) der "extrem niedrige" Bereich. Dieser macht statistisch (von der Normalverteilung gesehen) rund 2,3 % aus (in der Realität ist er um 1 bis 2 Prozentpunkte größer, weil sich in diesem Bereich die genetisch bedingten Minderbegabungen finden, die dafür verantwortlich sind, dass kognitive Fähigkeiten keine ganz symmetrischen Verteilungen aufweisen). Der obere Intelligenzbereich wird entsprechend aufgeteilt, **Tabelle 1** gibt die Übersicht dazu.

Tabelle 1: Interpretationsbereiche der IQ-Skala

Bereich	Anteil	Qualifizierung
unter 70	2,3	extrem niedrig
70-80	6,9	sehr niedrig
80-90	16,1	niedrig
90-110	49,5	durchschnittlich
110-120	16,1	hoch
120-130	6,9	sehr hoch
über 130	2,3	extrem hoch

Visualisierung der Testergebnisse

Die TDB2OnlineApp ist eine Internet-Applikation, zu deren besonderen Merkmalen eine einheitliche graphische Aufbereitung und Darstellung von Testergebnissen zählt. Die dazu verwendete Methode wird seit 2009 in der Klinik und Poliklinik für Psychiatrie und Psychotherapie des Klinikums der Universität München eingesetzt und hat die verbale Befundbeschreibung erleichtert und vereinheitlicht. Die Besonderheiten der Methode werden im Folgenden erläutert.

Einschätzung der absoluten Leistung

Das graphische Profil der TDB2OnlineApp zeichnet jeden Testwert als schwarzen Punkt in ein Leistungswertgitter ein, das von Test zu Test gleichbleibt und auf dem sich die Leistung über verschiedene Testverfahren hinweg vergleichen lässt. Die Skala mit der kleinen Beschriftung "Leistungswert", die dem Gitter zugrunde liegt, geht bei jedem Test von 40 bis 145 und die Werte stehen immer an der gleichen Position. Die Zahlen innerhalb des Rechtecks, das einen einzelnen Subtest darstellt, geben die Position der möglichen Rohwerte des Subtests an. In **Abbildung 2** lässt sich erkennen, dass ein Rohwert von 30 im WST ungefähr dem Leistungswert 100 entspricht, ein solcher von 20 ungefähr dem Leistungswert 85. Die Leistungswerte selbst sind von Test zu Test vergleichbar. Damit ist die Aussage möglich, dass bei diesem Patienten die Leistung im Lernen einer Wortliste (VLMT, Lerndurchgang 5, Rohwert 14, Leistungswert etwa 111) um rund 27 IQ-Punkte (oder knapp zwei Standardabweichungen) besser ist als seine Leistung im Teil B des Pfadfindertests (Rohwert 79, Leistungswert etwa 84). Ohne die gemeinsame Messebene der Leistungswerte wären solche Profilinterpretationen hinsichtlich der gezeigten Leistung nicht möglich.

Ablesen des Messbereichs

Der kleinste und der größte Rohwert, den ein Test liefern kann, ist immer im Profil eingezeichnet, wenn er im darstellbaren Leistungswertbereich zwischen 40 und 145 liegt. Man kann in **Abbildung 2** am Beispiel des WST also sehen, dass der kleinstmögliche Rohwert 1 und der größtmögliche Rohwert 41 ist. Das entspricht einem relativ großen Messbereich, weil die entsprechenden Leistungswerte von ungefähr 60 bis ungefähr 140 reichen. Beim Subtest *Verzögertes Wiedererkennen* im VLMT ist das anders. Dieser Subtest erreicht seine Testdecke (den größtmöglichen Rohwert von 15) schon bei einem Leistungswert von etwa 114. Allerdings würde der potenzielle Messbereich sehr weit nach unten reichen: Rohwerte unter 9 werden im Darstellungsbereich des Testprofils gar nicht mehr erfasst (und sind auch nicht mehr normiert, weil sie extrem selten vorkommen).

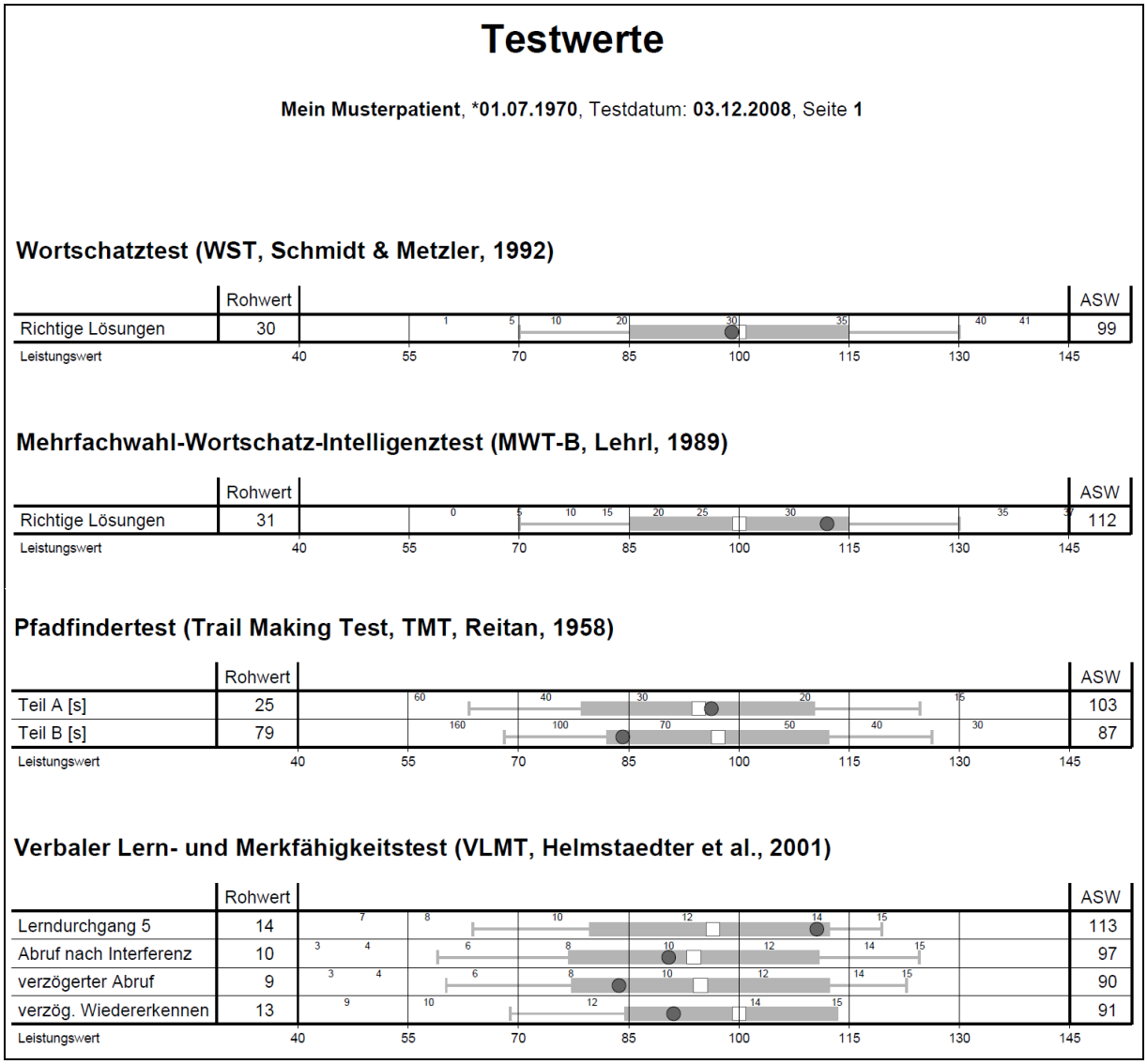


Abbildung 2: Beispielprofil für einen 38-jährigen Patienten

Bewertung der Messgenauigkeit

Eine Einschätzung der Messgenauigkeit erhält man, wenn man die Auflösung der Rohwertskala mit der Auflösung der Leistungswertskala vergleicht. Beim WST (siehe **Abbildung 2**) entsprechen die Rohwerte zwischen 10 und 20 ungefähr Leistungswerten zwischen 75 und 85. An dieser Stelle der Skala erhöht also ein zusätzliches richtiges Wort den IQ um einen Punkt. Dies ist gleichbedeutend mit einer relativ hohen Messgenauigkeit. Im oberen Leistungsbe-

reich (für den der Test auch nicht in erster Linie gedacht ist) erhöht ein zusätzliches richtiges Wort (z. B. von Rohwert 40 auf 41) den IQ um 6 oder 7 Punkte. Wenn ein einziges zusätzlich gewusstes (oder gar geratenes) richtiges Wort den IQ so stark verändern kann, ist die Messgenauigkeit an dieser Stelle der Skala ziemlich niedrig. An diesem Beispiel sieht man ganz praktisch, dass die Messgenauigkeit eines Tests keineswegs an allen Stellen gleich sein muss (wie die Reliabilitätstheorie und das von ihr abgeleitete Konzept des Standardmessfehlers suggerieren). Die meisten Tests messen im Mittelbereich genauer als an den Enden, eine Aussage, zu der man mit Hilfe der Item-Response-Theorie, einer Skalierungsmethode für Testitems, regelmäßig gelangt.

Ein Blick auf den Gedächtnistest VLMT (siehe **Abbildung 2**) reicht aus um zu erkennen, dass alle vier Subtests nur relativ grob messen: Ein einziges zusätzlich gelerntes oder erinnertes Wort macht viele IQ-Punkte auf der Leistungswertskala aus.

Der Leistungswert wird übrigens nicht numerisch ausgegeben, er lässt sich nur an der Position des schwarzen Punktes im Testprofil ablesen.

Vergleich mit der individuellen Altersnorm

Für die klinische Beurteilung eines Testwertes ist vor allem bei älteren Patienten ein Vergleich mit den Leistungen der entsprechenden Altersgruppe notwendig. Da die Werte im tdb2-Testprofil graphisch als altersunabhängige Leistungswerte erscheinen, muss man die Vergleichsbereiche der Altersgruppe zusätzlich einzeichnen. Als Vergleichswerte böten sich zunächst der Mittelwert der Altersgruppe und deren Standardabweichung an. Leider sind viele Verteilungen psychologischer Testwerte nicht normalverteilt (entweder weil die Verteilung selbst schief ist oder weil der Messbereich irgendwo willkürlich endet und die Verteilung dadurch gekappt ist). Beim Einzeichnen von Mittelwerten und Ein- oder Zwei-Sigma-Grenzen würden zumindest die Zwei-Sigma-Grenzen oft weit über den verfügbaren Messbereich hinausgehen und damit sinnlose Grenzwerte anzeigen. In tdb2 werden deshalb Vergleichsbereiche eingezeichnet, die über Prozentränge (PR) ermittelt wurden. Das weiße Quadrat kennzeichnet den Median (= PR 50) der Altersgruppe des Patienten, das graue Rechteck den Bereich von PR 16 bis PR 84 und die beiden Fähnchen an den Enden des grauen Rechtecks die Prozentränge 2,5 und 97,5. Wenn die Testwerte normalverteilt sind, entsprechen die genannten Prozentränge genau dem Mittelwert und den Ein- und Zwei-Sigma-Grenzen.

Man kann die individuelle Stellung eines Patienten in den Grenzen seiner Altersbezugsgruppe natürlich aus der Grafik ablesen, allerdings ist das etwas mühsam. Deshalb werden die altersnormierten Standardwerte vom Programm ausgerechnet und auf der rechten Seite des Profils in der Spalte "ASW" angezeigt. Auch diese altersnormierten Standardwerte werden IQ-skaliert angezeigt, also mit Mittelwert 100 und Standardabweichung 15. An einem Beispiel kann man zeigen, wie der Wert berechnet wird: In der vorletzten Zeile der **Abbildung 2** (*Verzögerter Abruf* im VLMT) kann man erkennen, dass der individuelle Testwert des Patienten um etwa zwei Drittel des grauen Bereichs vom weißen Quadrat nach links entfernt liegt. Wäre er auf dem weißen Quadrat, wäre das der altersnormierte Standardwert 100. Wäre er am linken Ende des grauen Bereichs, wäre das der altersnormierte Standardwert 85. Da er tatsächlich etwa um 2/3 des grauen Bereichs unterhalb des weißen Quadrats liegt, entspricht dies dem altersnormierten Standardwert 90. Dieser Wert ist in der rechten Spalte eingetragen.

Bei jungen Erwachsenen weichen die altersbezogenen Standardwerte kaum von den Leistungswerten ab. Je älter ein Patient ist und je mehr die Leistung in einem bestimmten Test mit zunehmendem Alter abnimmt, umso größer ist der Unterschied zwischen Leistungswerten und altersbezogenen Standardwerten. Man sieht das deutlich in **Abbildung 3**, die die gleichen Testwerte zeigt wie **Abbildung 2**, aber dieses Mal für einen Patienten im Alter von 78 statt 38 Jahren. Die Position der schwarzen Punkte im Leistungswertgitter bleibt gleich. Was sich ändert, sind die Altersreferenzbereiche und damit die altersbezogenen Standardwerte in der

Spalte ASW. Die Veränderung des Alters des Patienten wirkt sich sehr stark aus beim visokonstruktiven Test TMT und beim Gedächtnistest VLMT. Sie hat keinen Effekt bei den beiden Wortschatztests, weil es bei beiden keine Normen für alte Personen gibt. Beim WST umfasst der Normbereich noch das Alter 78, beim MWT-B nicht mehr. TDB2Online verwendet in solchen Fällen den nächst passenden Altersreferenzbereich, fügt aber eine entsprechende Bemerkung an. Nach den Daten anderer Wortschatztests wäre übrigens mit einem gewissen Altersabbau zu rechnen, wenn es hier vernünftige Normdaten gäbe. Der Effekt wäre aber viel geringer als bei TMT und VLMT.

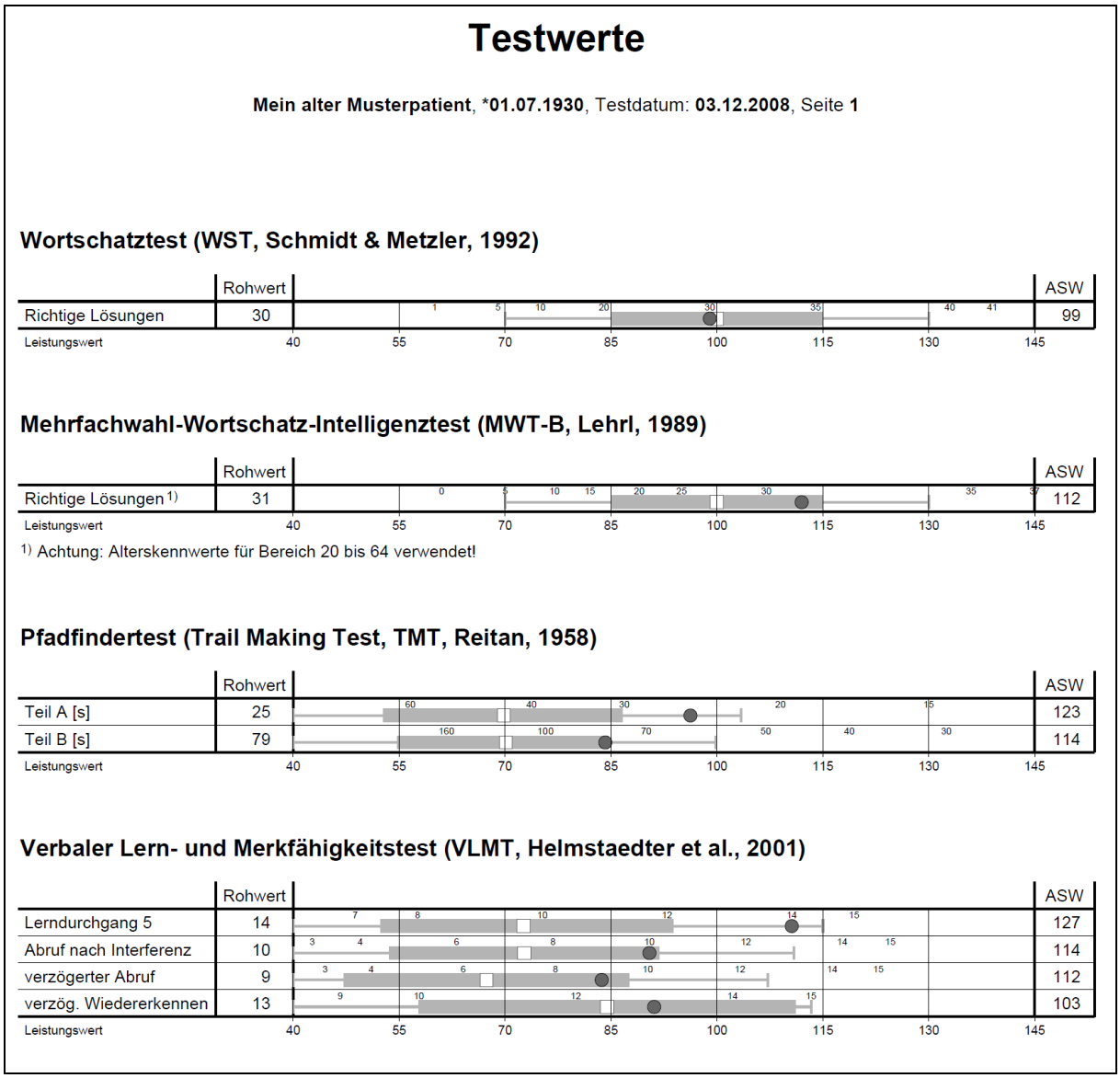


Abbildung 3: Testleistungen wie Abbildung 2 bei einem um 40 Jahre älteren Patienten

Vorgehen bei Globalwerten (z. B. Gesamt-IQ)

Wie bei vielen anderen Intelligenztestbatterien ist es auch bei den Wechsler-Tests üblich, einen Indexwert für einige ("Verbal-IQ", "Handlungs-IQ") bzw. alle Subtests ("Gesamt-IQ") zu bilden. Die Praxis geht auf den Beginn der Intelligenztestdiagnostik zurück, als vor allem die globale intellektuelle Begabung erfasst werden sollte und weniger deren Struktur. In der aktuellen neuropsychologischen Diagnostik liegt der Fokus dagegen viel stärker auf der differenzierten Erfassung von Einzelleistungen und weniger auf pauschalen Begabungskennwerten.

Die gleichzeitige Erfassung von Einzelleistungen und Globalwerten hat nun gewisse psychometrische Tücken. An einem Beispiel kann man das schnell erklären. Nehmen wir an, zwei Einzelleistungen (Allgemeinwissen und Kopfrechnen) sollen auf einer IQ-Skala mit Mittelwert 100 und Standardabweichung 15 sowohl einzeln erfasst und dargestellt als auch zu einem "Verbal-IQ" kombiniert werden. Als psychometrisch naiver Betrachter denkt man vermutlich, dies sei einfach und der Verbal-IQ errechne sich als Mittelwert der beiden Einzelleistungen. Dem ist aber keineswegs so, weil in den herkömmlichen IQ-Tests die Einzelleistungen und die Gesamtwerte getrennt standardisiert werden. Die doppelte Standardisierung führt dazu, dass die Korrelation zwischen den beiden Subtests darüber entscheidet, ob für die Kombination die gleiche oder eine andere Metrik verwendet wird. Je niedriger die Korrelation zwischen den Subtests ist (und je größer die Anzahl der einbezogenen Subtests ist) desto mehr weicht die neue Metrik von der alten ab. **Tabelle 2** zeigt die Konsequenzen an unserem einfachen Beispiel. Wenn beide Subtestwerte 100 betragen und damit genau in der Mitte der Verteilung liegen, spielt die Höhe der Korrelation keine Rolle. In diesem Fall beträgt der kombinierte "Verbal-IQ" immer 100. Anders wird es, wenn die Einzelwerte von 100 abweichen. Betragen beide Einzelwerte genau 85 IQ-Punkte, dann wäre der "Verbal-IQ" nur dann auch 85, wenn die beiden Einzelleistungen mit 1 miteinander korrelieren. In der Praxis liegen die Korrelationen zwischen Subtests, die zu einem Gesamtwert verrechnet werden, meist irgendwo zwischen .30 und .70. Bei der schon relativ hohen Korrelation von .70 sinkt der "Verbal-IQ" auf 84 statt 85 ab, bei der eher niedrigen von .30 schon auf 81 statt 85. Je extremer sich die Einzelwerte vom Mittelwert der Verteilung entfernen desto größer werden auch die Unterschiede. Liegen die Einzeltestwerte bei 70 IQ-Punkten, dann verringert sich der "Verbal-IQ" bei einer Korrelation von .70 auf 66 Punkte, bei einer Korrelation von .30 auf 63 Punkte. Der "Spreizfaktor" für die Metrik wird umso größer je niedriger die Korrelation zwischen den Subtests ist und je mehr Subtests kombiniert werden.

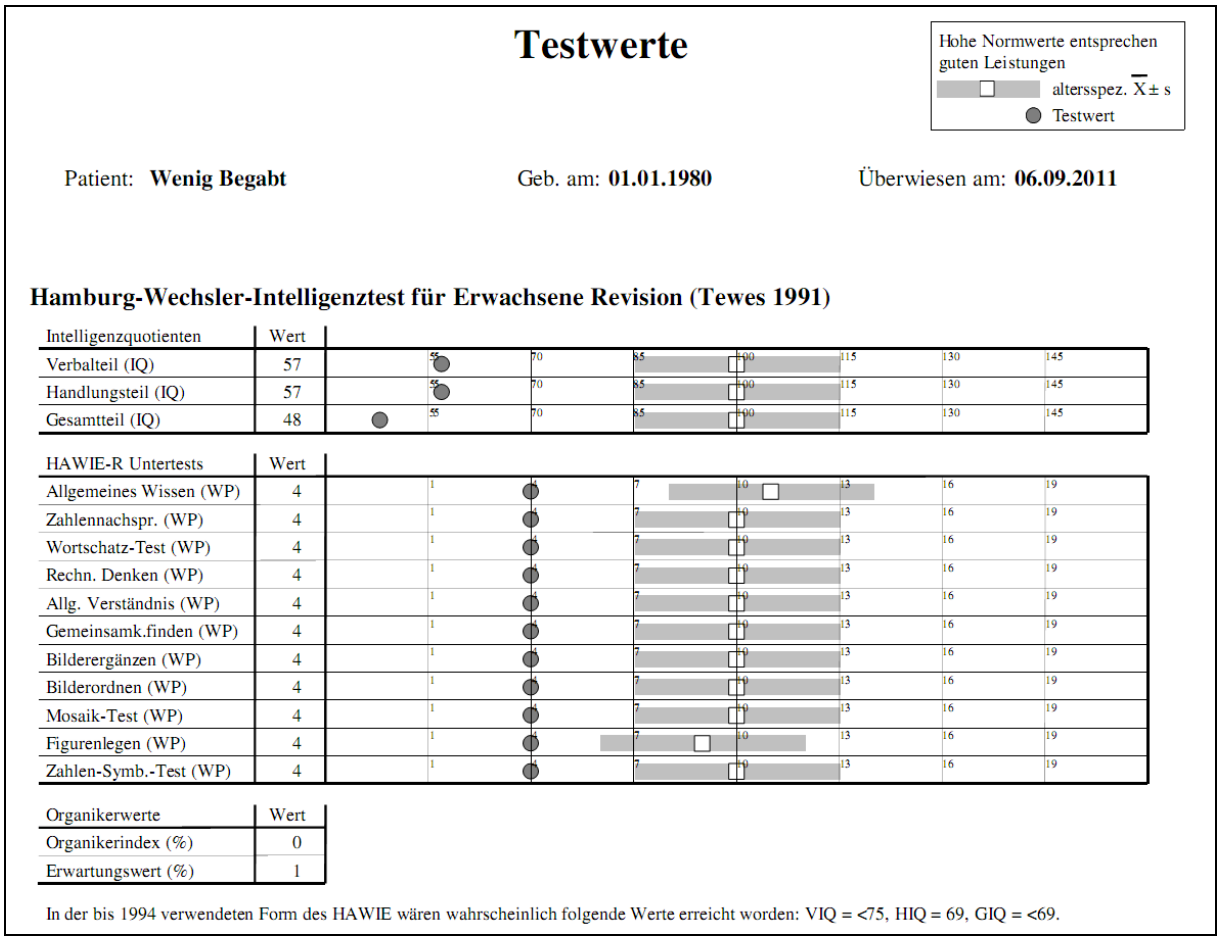
Tabelle 2: Auswirkungen einer getrennten Standardisierung auf die Metrik

Einzelleistungen	Korrelation	"Verbal-IQ"
100	1	100
	0	100
85	1	85
	.7	84
	.3	81
	0	79
70	1	70
	.7	66
	.3	63
	0	58
130	1	130
	.7	134
	.3	137
	0	142

Obwohl diese psychometrischen Beziehungen zwischen Subtests und Globalmaßen eigentlich bekannt sind, wurden sie in der klinischen Diagnostik praktisch nicht beachtet. Das änderte sich im Erwachsenenbereich mit der Einführung des HAWIE-R als Nachfolger des HAWIE. Damals wunderten sich viele Praktiker darüber, dass speziell bei niedrig Begabten die mit dem neuen HAWIE-R berechneten IQs viel niedriger waren als die zuvor mit dem HAWIE bestimmten. Dies hing zum einen mit dem Phänomen des IQ-Zugewinns der Bevölkerung

über die Zeit hinweg zusammen ("Flynn-Effekt"), das dazu führt, dass neuere Verfahren niedrigere IQs messen als früher standardisierte. Allerdings erklärte das nur einen Teil der Differenz. Der andere Teil kam daher, dass die mittlere Interkorrelation der Subtests in der Standardisierungsstichprobe des HAWIE-R (warum auch immer) lediglich .40 betrug, in den für die IQ-Berechnung benutzten Altersklassen 20-34 Jahre sogar nur .32, beides deutlich niedriger als beim alten HAWIE, wo die mittlere Subtestinterkorrelation etwa bei .60 lag. Die niedrigere Interkorrelation der Subtests in der Standardisierungsstichprobe führte beim HAWIE-R zu einer größeren Spreizung der Metrik der IQs im Verhältnis zur Metrik der Subtests.

Abbildung 4 zeigt das an einem Beispielfall eines 31-jährigen Probanden mit niedriger Begabung, der in sämtlichen Subtests des HAWIE-R lediglich vier Wertpunkte erzielte. Die klassischen, nicht altersabhängigen Wertpunkte des HAWIE-R haben einen Mittelwert von 10 und eine Standardabweichung von 3. Vier Wertpunkte entsprechen deshalb einem Wert von 70 auf der IQ-Skala. Wegen der niedrigen Interkorrelation der Subtests betragen aber die tatsächlich bestimmten Teil-IQs statt der naiv erwarteten 70 nur 57 für den Verbalteil und 57 für den Handlungsteil. Weil der Effekt auf die Testmetrik umso größer wird, je mehr Subtests einbezogen werden, beträgt der Gesamt-IQ nur 48. Auf Grund der Subtestergebnisse hätte man einen von 70 erwartet, nach den Teil-IQs einen von 57. (**Abbildung 4** wurde übrigens mit einem alten, von 1995 bis 2008 verwendeten Vorgängerprogramm („tdb“) erzeugt, das die nach Handbuchvorschrift berechneten Wertpunkte und IQs darstellte und lediglich den altersspezifisch erwarteten Normbereich zusätzlich einzeichnete.)



Testwerte

Hohe Normwerte entsprechen guten Leistungen
altersspez. $\bar{X} \pm s$
Testwert

Patient: **Wenig Begabt** Geb. am: **01.01.1980** Überwiesen am: **06.09.2011**

Hamburg-Wechsler-Intelligenztest für Erwachsene Revision (Tewes 1991)

Intelligenzquotienten	Wert
Verbalteil (IQ)	57
Handlungsteil (IQ)	57
Gesamtteil (IQ)	48

HAWIE-R Untertests	Wert
Allgemeines Wissen (WP)	4
Zahlennachspr. (WP)	4
Wortschatz-Test (WP)	4
Rechn. Denken (WP)	4
Allg. Verständnis (WP)	4
Gemeinsamk.finden (WP)	4
Bilderergänzen (WP)	4
Bilderordnen (WP)	4
Mosaik-Test (WP)	4
Figurenlegen (WP)	4
Zahlen-Symb.-Test (WP)	4

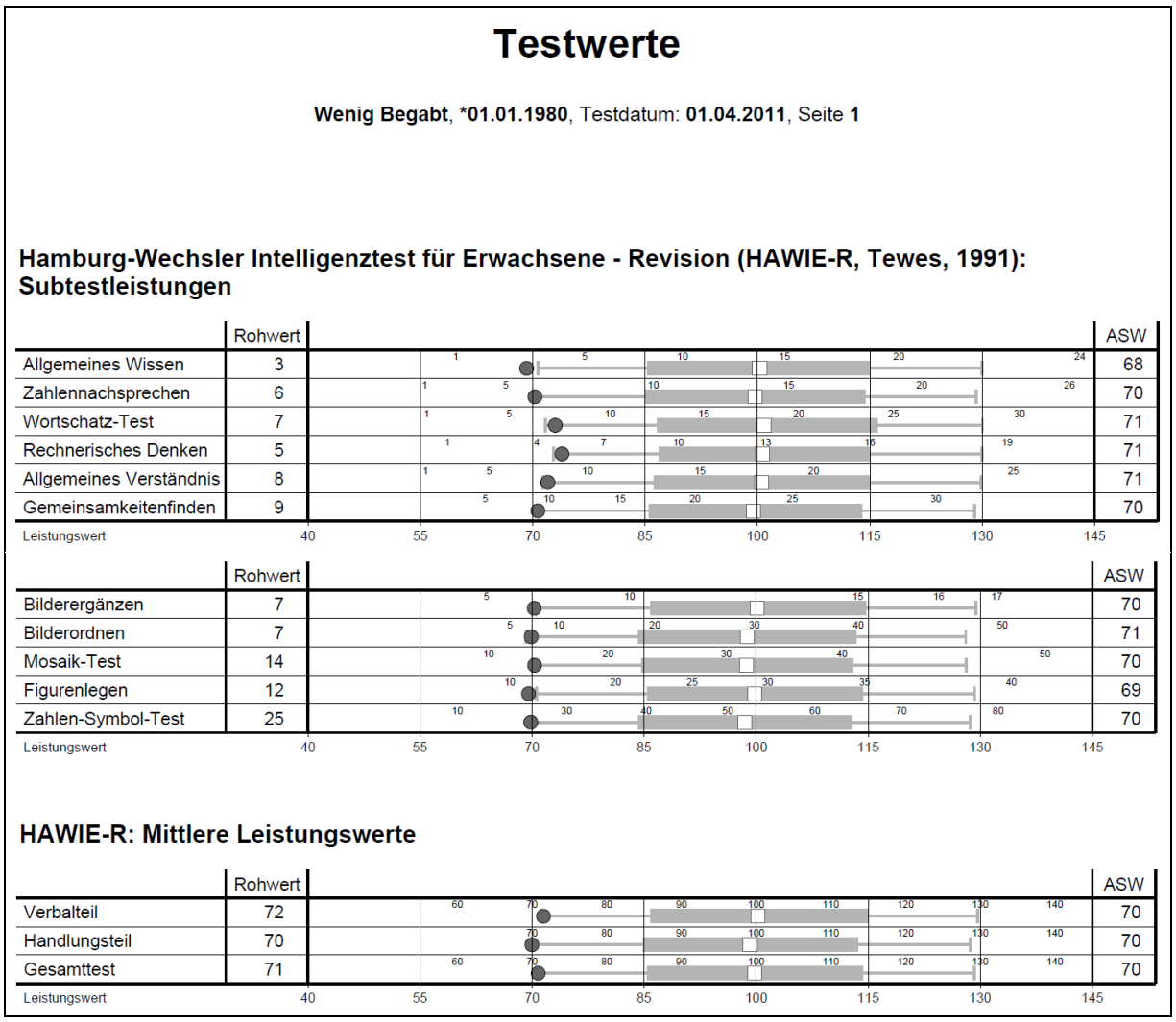
Organikerwerte	Wert
Organikerindex (%)	0
Erwartungswert (%)	1

In der bis 1994 verwendeten Form des HAWIE wären wahrscheinlich folgende Werte erreicht worden: VIQ = <75, HIQ = 69, GIQ = <69.

Abbildung 4: Beispiel für die unterschiedliche Metrik von Subtests und IQs beim HAWIE-R (Ausgabebeispiel von tdb, klassische Wertpunkte, IQs nach Handbuch berechnet)

Niemand kann in der Praxis mit solch unterschiedlichen Metriken wirklich umgehen. Die Idee, dass zu unterschiedlichen Aggregationsgraden der Tests und Subtests auch unterschied-

liche Metriken gehören, ist psychometrisch stimmig, praktisch dagegen kaum zu vermitteln. Dass viele "niedrige" Einzelleistungen zu einem "sehr niedrigen" Gesamtergebnis führen, ist für die meisten Beurteiler unverständlich. In der Notengebung zum Beispiel wäre das nicht denkbar, dort wird gemittelt und vier Einzelnoten von "ausreichend" führen auch zu einer Gesamtnote von "ausreichend". Wollte man die psychometrisch richtige Aggregation unter Berücksichtigung der unterschiedlichen Korrelationen der Subtests tatsächlich beibehalten, müsste man die verbalen Umschreibungen der Testergebnisse an die Testebenen anpassen. Dies wäre extrem schwierig, weil die Verhältnisse von Test zu Test und von Normierung zu Normierung unterschiedlich sind.



Testwerte

Wenig Begabt, ***01.01.1980**, Testdatum: **01.04.2011**, Seite **1**

Hamburg-Wechsler Intelligenztest für Erwachsene - Revision (HAWIE-R, Tewes, 1991): Subtestleistungen

	Rohwert	ASW
Allgemeines Wissen	3	68
Zahlennachsprechen	6	70
Wortschatz-Test	7	71
Rechnerisches Denken	5	71
Allgemeines Verständnis	8	71
Gemeinsamkeitenfinden	9	70

Leistungswert 40 55 70 85 100 115 130 145

	Rohwert	ASW
Bilderergänzen	7	70
Bilderordnen	7	71
Mosaik-Test	14	70
Figurenlegen	12	69
Zahlen-Symbol-Test	25	70

Leistungswert 40 55 70 85 100 115 130 145

HAWIE-R: Mittlere Leistungswerte

	Rohwert	ASW
Verbalteil	72	70
Handlungsteil	70	70
Gesamttest	71	70

Leistungswert 40 55 70 85 100 115 130 145

Abbildung 5: Mittlere Leistungswerte als Ersatz für die IQs (Beispiel aus tdb2, nur Subtestrohwerte werden eingegeben)

In einem neuropsychologischen Kontext ist es viel sinnvoller, durchgehend nur eine Metrik zu benutzen, und zwar die der Subtestebene. Nur dort findet die differenzierte Erfassung unterschiedlicher Fähigkeiten statt. Deshalb wird in tdb2 (und damit in TDB2Online) jede Einzelleistung mit einer über die Tests hinweg vergleichbaren Metrik dargestellt. Globalwerte mit eigener, anderer Standardisierung werden ins Profil nicht aufgenommen. Stattdessen werden die zu einem Globalwert gehörenden Leistungswerte der Einzelverfahren schlicht gemittelt. **Abbildung 5** zeigt an einem Beispielfall, dass die gleichen Testleistungen wie in **Abbildung 4** in tdb2 zu sogenannten "mittleren Leistungswerten" agglutiniert werden, die auf der gleichen Metrik liegen wie die Einzelleistungen. Wie sonst auch werden zusätzlich die Al-

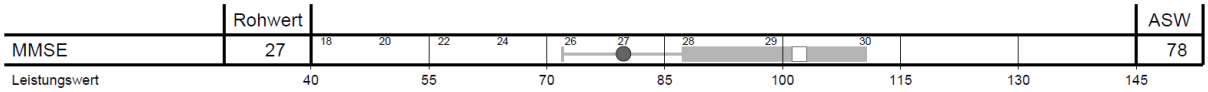
tersnormgrenzen eingeblendet und ein alterskorrigierter Standardwert (ASW) für den Aggregatwert berechnet und in der rechten Spalte des Profils dargestellt. Diese alterskorrigierten Standardwerte entsprechen in ihrer Zusammensetzung den üblichen Intelligenzquotienten (sie geben die durchschnittliche Fähigkeit über die einbezogenen Subtests an und sind altersstandardisiert), sind aber in der gleichen Metrik ausgedrückt wie die Subtests. Sprachlich grenzen wir sie von den (separat standardisierten) Intelligenzquotienten des HAWIE-R, WIE und WAIS-IV bzw. von den Indexwerten des WIE und des WAIS-IV dadurch ab, dass wir von mittleren alterskorrigierten Standardwerten (100;15) sprechen.

Besonderheiten bei den Verfahren für die Demenzdiagnostik

Bei den meisten Testverfahren, die für die Demenzdiagnostik verwendet werden, gibt es keine Normen für junge Erwachsene. Ein Beispiel für solche Tests ist die CERAD-Testbatterie. Bei diesem und anderen Verfahren gibt es Normen für gesunde Probanden nur im Altersbereich von etwa 50 bis 90 mit einem Mittelwert um die 70 Jahre, oft nicht sonderlich gut spezifiziert. Bei solchen Tests beziehen sich die Leistungswerte nicht auf junge Erwachsene, sondern auf Personen von etwa 70 Jahren. In der Überschrift zu den jeweiligen Testverfahren ist die Quelle der Leistungsnormdaten immer dann angegeben, wenn sie vom Standard (junge Erwachsene) abweicht (siehe **Abbildung 6**).

Abbildung 6 zeigt noch eine weitere Besonderheit der Demenztests. Diese Tests sind nicht zur Quantifizierung von Fähigkeiten über einen breiten Normbereich hinweg konstruiert. Bei ihnen steht der Nachweis und die Quantifizierung von Defiziten im Vordergrund. In fast allen Demenztests erreicht ein Gesunder schon mit mittleren Leistungen die Testdecke. Am auffälligsten ist das beim Subtest *Wörter Wiedererkennen* in der CERAD. In diesem Subtest erreichen mehr als 50 Prozent der Normstichprobe der 70-Jährigen die volle Punktzahl von 10 Richtigen. Man sieht das daran, dass der maximal erreichbare Rohwert von 10 zu einem Leistungswert von unter 100 führt. In solchen Fällen geben wir in der Spalte für die altersbezogenen Standardwerte statt eines festen Wertes nur noch einen Bereich an, hier also ">95".

Mini Mental State Examination (Folstein et al., 1975) - Normbasis: gesunde Personen um die 70



CERAD Screening-Batterie für Alzheimer Demenz (Welsh et al., 1994) - Normbasis: gesunde Personen um die 70

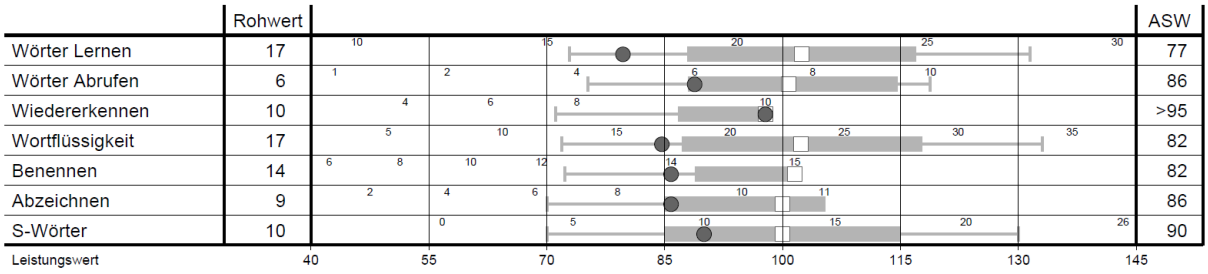


Abbildung 6: Angabe der Normbasis im Titel bei Abweichungen vom Standard