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

Darstellung von Testergebnissen in TDB2Online

Revisionsstand: 13.10.2026

In knapper Form werden die Konzepte und Verfahren von TDB2Online erläutert, insbesondere die Eigenschaften der gewählten Visualisierung. Die ausführliche Dokumentation steht auf der TDB2Online-Webseite, auf die gelegentlich verwiesen wird. Die speziellen Dokumentationen zu den Testverfahren sind nur für angemeldete professionelle Nutzer zugänglich.

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 (Allgemeines Wissen in der WAIS)
- Anzahl der Fehler in einer Kategorisierungsaufgabe (Halstead Category Test)
- Zeit in Sekunden für die Durchführung einer visokonstruktiven Aufgabe (Trail Making Tests)
- Anzahl richtig reproduzierter Wörter in einem Gedächtnistest (Wortliste der CERAD)
- Differenz zwischen der Anzahl richtig wiedererkannter Wörter minus Anzahl der fälschlich "wiedererkannten" Wörter in einer Wiedererkennensaufgabe (Wörter Wiedererkennen in der CERAD)
- Punkte in einer komplexen Aufgabe, die nach Richtigkeit und Schnelligkeit der Durchführung bewertet wird (Mosaiktest in der WAIS)

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 (LW) 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 Standardisierung von Leistungswerten erfolgt idealerweise 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 (ASW)**. 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 der Fähigkeitsmessung verlassen und die Ebene der Interpretation betreten wird. Mehr dazu hier.

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 anhand 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 Wert von 130 auf der IQ-Skala einem Prozentrang von 97,7 entspricht, was nichts anderes bedeutet, als dass 97,7 Prozent der Vergleichsstichprobe einen niedrigeren Wert als 130 haben und die restlichen 2,3 Prozent einen höheren. Zu jedem Wert auf der IQ-Skala gibt es einen entsprechenden Prozentrangwert, den man ausführlicheren Tabellen entnehmen kann.

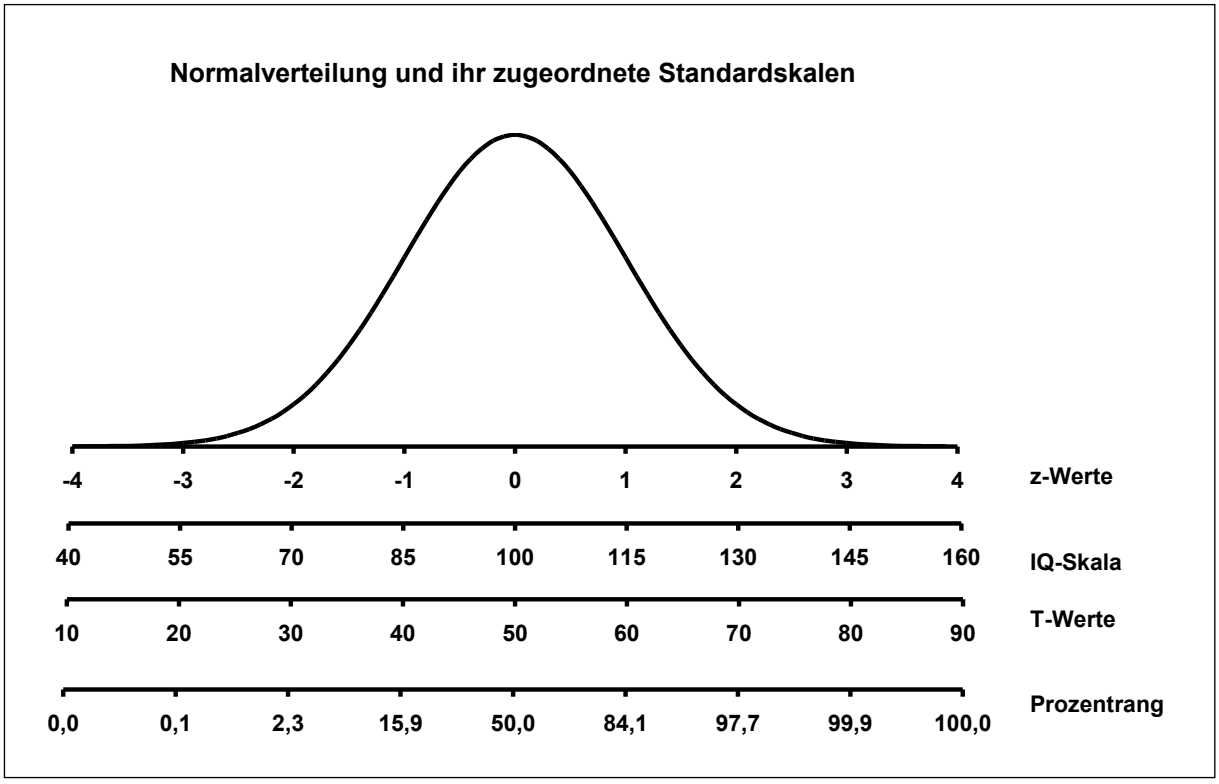


Abbildung 1: Normalverteilung und einige der von ihr abgeleiteten Standardskalen

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 rund 2,3 % aus. 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

Bei manchen Tests sind die Rohwerte leider nicht so normal verteilt, dass obige Annahmen und Prozentzahlen auf sie zutreffen. Meistens muss man die Skala dann linearisieren. Mehr dazu in der ausführlichen Doku unter dem Stichwort Linearitätsannahme.

Visualisierung der Testergebnisse

Die TDB2OnlineApp ist eine Web-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. Sie vermittelt Informationen, die über die Angabe eines einzelnen Standardwerts weit hinausgehen und hat die verbale Befundbeschreibung erleichtert und vereinheitlicht. Die Besonderheiten werden im Folgenden erläutert.

Referenzgruppe für die Leistungswerte

Die ideale Referenzgruppe für die Erstellung von Leistungswerten wäre, wie oben schon gesagt, eine große Stichprobe von jungen Erwachsenen im Alter von 20-24 Jahren. Leider steht eine solche Gruppe nicht bei allen Testverfahren zur Verfügung. Bei den neuropsychologischen Testverfahren, die speziell für den Demenzbereich entwickelt wurden, also zum Beispiel der CERAD-Batterie, gibt es keine Normdaten für junge Erwachsene. Aber auch bei einigen anderen Tests ist keine ideale Referenzgruppe verfügbar. Das TDB2Online-Testprofil zeigt deshalb im Titel des Tests jeweils die Altersgruppe, die als Referenz für die Leistungswerte dient. Für die WAIS-IV heißt das in **Abbildung 2** zum Beispiel: LW-Basis: 20-29.

Einschätzung der absoluten Leistung

Im Testprofil der TDB2OnlineApp wird jeder Testwert als schwarzer Punkt auf einer Leistungswertskala eingetragen, die von Test zu Test gleichbleibt und auf der sich die Leistung über verschiedene Testverfahren hinweg vergleichen lässt. Die Skala mit der kleinen Beschriftung "Leistungswert" geht bei jedem Test von 40 bis 145 und die Werte stehen immer an der gleichen Position. Die senkrechten dünnen Striche markieren jeweils eine Standardabweichung. Die Zahlen innerhalb des Rechtecks, das einen einzelnen Subtest darstellt, geben die Position der Rohwerte des Subtests an. In **Abbildung 2** lässt sich erkennen, dass ein

Rohwert von 30 im Wortschatztest 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 24-jährigen Patienten die Erinnerungsleistung beim Abrufen einer Wortliste (VLMT, *Abruf nach Interferenz*, Rohwert 12, Leistungswert etwa 104) um rund 20 IQ-Punkte (mehr als eine Standardabweichung) besser ist als seine Leistung im Teil B des TMT (Rohwert 79, Leistungswert etwa 84). Ohne die gemeinsame Messebene der Leistungswerte wären vergleichende Interpretationen der Leistungen in unterschiedlichen Domänen nicht möglich.

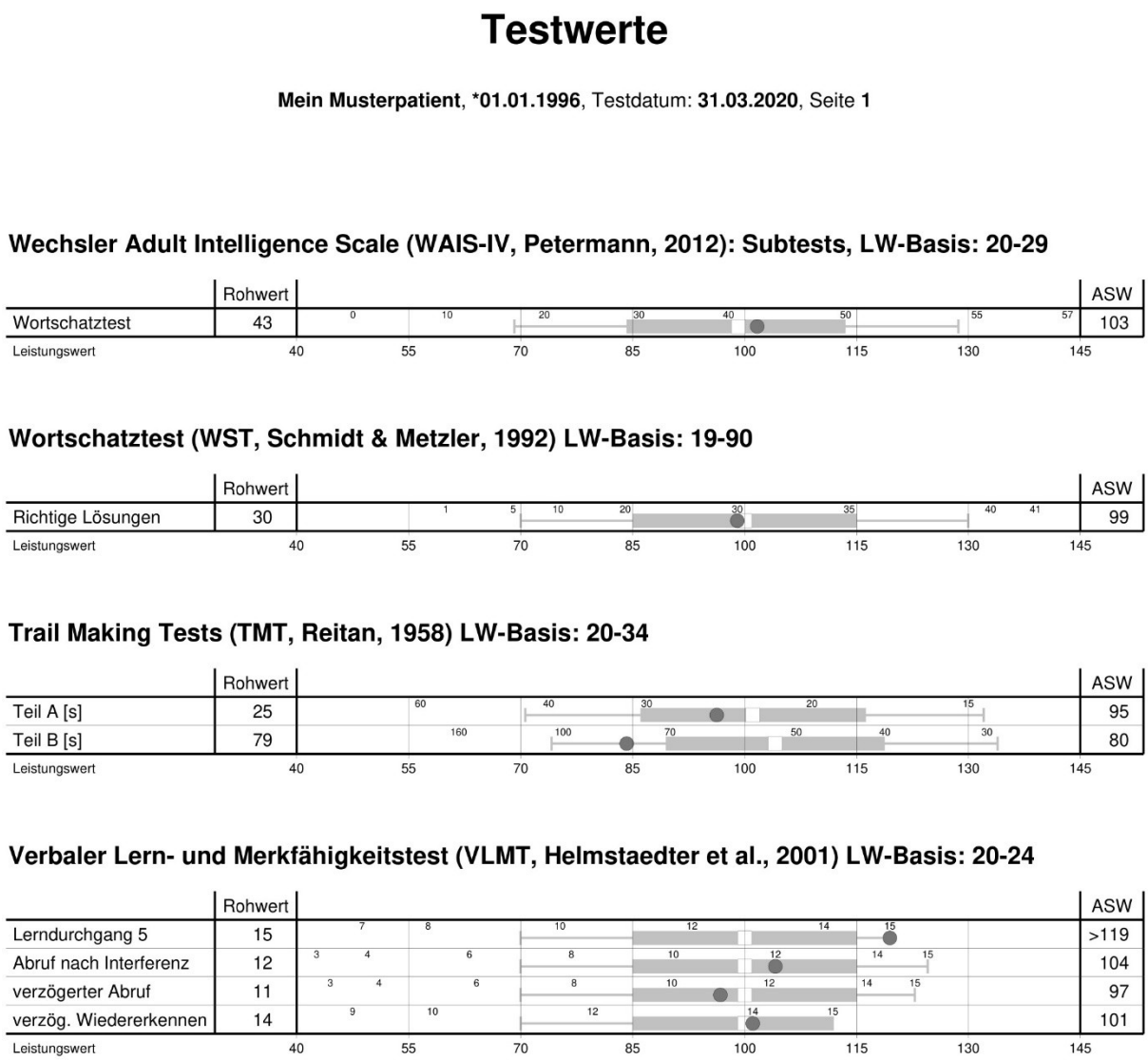


Abbildung 2: Beispielprofil für einen 24-jährigen Patienten

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 Wortschatztests WST also nachverfolgen, 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 113. 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).

Bewertung der Messgenauigkeit

Eine erste Einschätzung der Messgenauigkeit erhält man, wenn man die Auflösung der Rohwertskala mit der Auflösung der Leistungswertskala vergleicht. Beim Wortschatztest 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 Leistungswert um einen Punkt. Dies bedeutet eine hohe Messgenauigkeit. Im oberen Leistungsbereich (für den der Test auch nicht in erster Linie gedacht ist) erhöht ein zusätzliches richtiges Wort (z. B. von Rohwert 40 auf 41) den Leistungswert um 6 oder 7 Punkte. Wenn ein einziges zusätzlich gewusstes (oder gar geratenes) richtiges Wort den Leistungswert 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 ist (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 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 ein Vergleich mit den Leistungen der entsprechenden Altersgruppe notwendig, vor allem bei älteren Patienten. Da die Werte im Testprofil graphisch altersunabhängig als 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 Rohwerte psychologischer Tests 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 sinnlose Grenzwerte anzeigen. Im TDB2Online-Profil 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 Leistungswerte normalverteilt sind, was oft durch eine Linearisierung der Rohwerte erreicht werden kann, entsprechen die genannten Prozentränge dem Mittelwert und den Ein- und Zwei-Sigma-Grenzen.

Man kann die individuelle Stellung eines Patienten in den Grenzen seiner Altersbezugsgruppe aus der Grafik ablesen, indem man schaut, wo der schwarze Punkt im Verhältnis zu dem grauen Rechteck und den äußeren Fähnchen liegt. Das wäre allerdings 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 ersten Zeile des Trail Making Tests auf der **Abbildung 2** (*Teil A*) kann man erkennen, dass der individuelle Testwert des Patienten um etwa ein Drittel des grauen Bereichs links 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 1/3 des grauen Bereichs unterhalb des weißen Quadrats liegt, entspricht dies dem altersnormierten Standardwert 95. Dieser Wert ist in der rechten *ASW*-Spalte eingetragen.

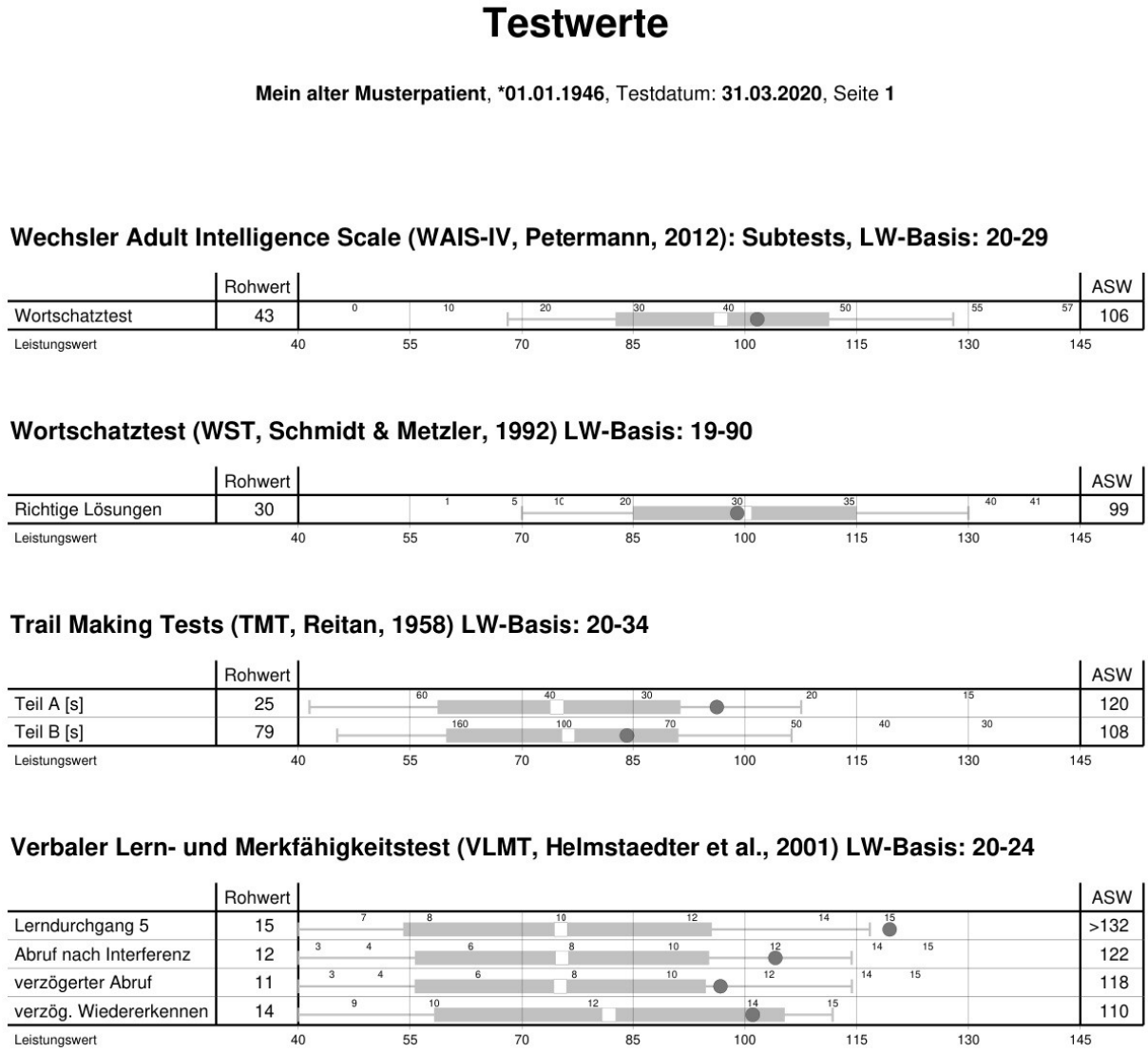


Abbildung 3: Testleistungen wie in Abbildung 2 bei einem 74 Jahre alten Patienten

Bei jungen Erwachsenen weichen die altersbezogenen Standardwerte kaum von den Leistungswerten ab. Bei normalverteilten Rohwerten reicht der graue Bereich im Allgemeinen von 85 bis 115, die Fähnchen nach unten bis 70 und nach oben bis 130. 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 in **Abbildung 3**, die die gleichen Testwerte zeigt wie **Abbildung 2**, aber dieses Mal für einen Patienten im Alter von 74 statt 24 Jahren. Die Position der schwarzen Punkte im Leistungswertgitter bleibt gleich. Was sich ändert, sind die ins Profil eingezeichneten Altersreferenzbereiche und dementsprechend die altersbezogenen Standardwerte in der Spalte ASW. Das veränderte Alter des Patienten wirkt sich sehr stark aus beim visokonstruktiven Trail Making Test (TMT, hier wird aus dem ASW von **95** beim Teil A in Abb. 2 ein ASW von **120** in Abb. 3) und beim Gedächtnistest VLMT. Sie hat keinen Effekt beim

Wortschatztest WST, weil es für den keine altersgestaffelten Normen gibt und das Alter von 74 Jahren noch im Altersbereich der Normstichprobe liegt. Im Wortschatztest der WAIS-IV sieht man etwas genauer, dass es im Alter von 74 Jahren auch im Wortschatz zu geringen Alterseinbußen in der Normstichprobe kommt, die unserem Probanden bei seiner gleichgebliebenen Leistung zu einem etwas höheren ASW von 106 verhelfen. Man kann aber schnell sehen, dass die Alterseffekte bei TMT und VLMT viel größer sind.

Grenzen des Messbereichs

Nicht alle Tests überspannen einen großen oder auch nur einen symmetrischen Leistungsbereich. Vorbildlich ist der Messbereich im Subtest *Allgemeines Wissen* in der WAIS-IV, der ungefähr von den Leistungswerten 47 bis 143 reicht (siehe **Abbildung 2 oder 3**). Mit den Subtests des VLMT lassen sich dagegen sehr hohe Leistungen nicht mehr messen. Bei allen Subtests ist 15 der höchst erreichbare Rohwert, er entspricht Leistungswerten zwischen 112 und 124, je nach Subtest. Beim ersten Subtest, dem Lernen der Wörter, im Profil von **Abbildung 2 und 3:** *Lerndurchgang 5*, führt ein Rohwert von 15 zu einem Leistungswert von etwa 120. Bei einem 24-Jährigen (**Abbildung 2**) entspricht das auch dem Altersstandardwert, bei einem 74-Jährigen (**Abbildung 3**) liegt der ASW dagegen etwa bei 133. In beiden Fällen ist mit Rohwert 15 aber das Ende des Messbereichs erreicht. Damit man sieht, dass höhere Leistungen nicht messbar sind, werden in diesen Fällen die Altersstandardwerte als nach oben offene Bereiche angegeben (>119 bzw. >132 im Beispiel). Analog wird verfahren, wenn der Messbereich eines Tests am unteren Ende limitiert ist.

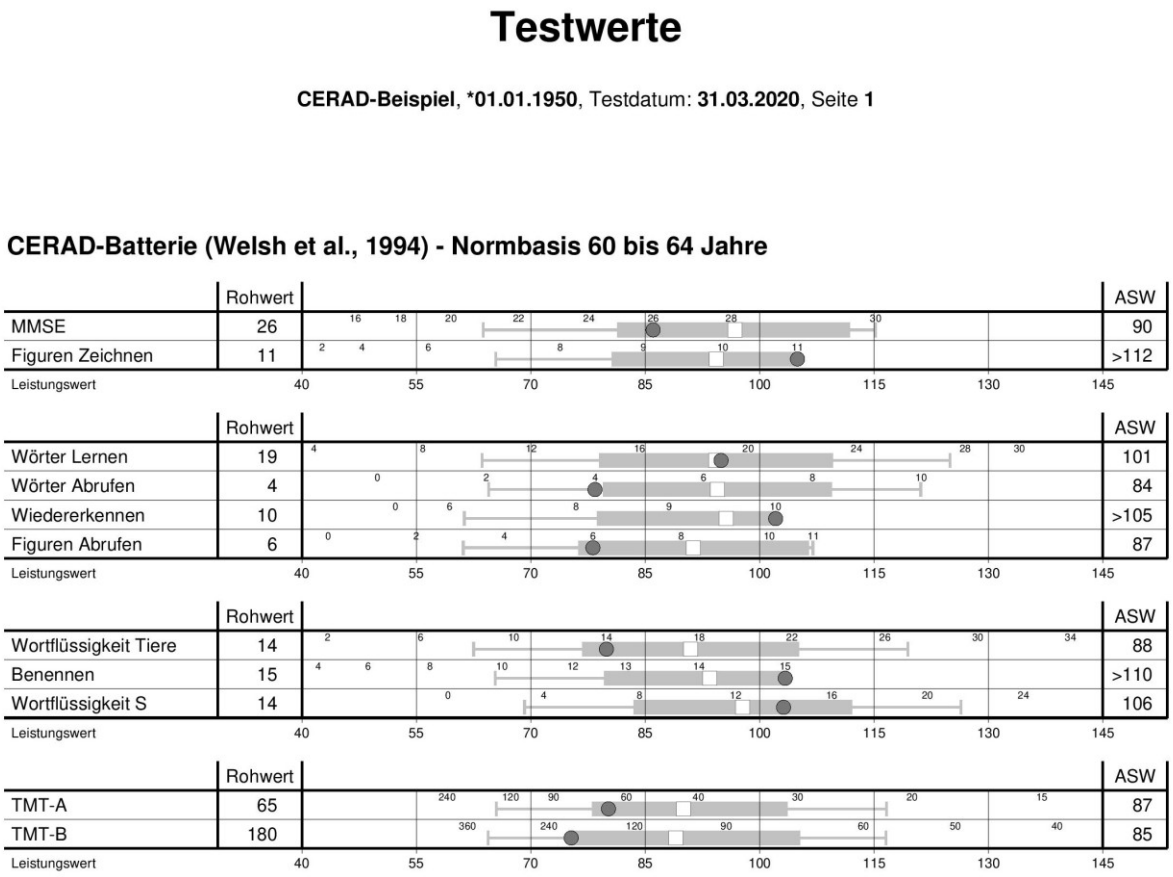


Abbildung 4: CERAD-Testwerte bei einem 70-Jährigen

Besonderheiten bei den Verfahren für die Demenzdiagnostik

Dass es bei vielen Testverfahren für die Demenzdiagnostik keine Normen für junge Erwachsene gibt, wurde oben schon gesagt. Ein prominentes Beispiel für solche Tests ist die CERAD-Testbatterie. In **Abbildung 4** sind CERAD-Testergebnisse für einen siebzigjährigen

Patienten dargestellt. Auch hier fallen – ähnlich wie zuvor beim VLMT in den Abbildungen 2 und 3 – die eingeschränkten Messbereiche und die dadurch extrem schiefen Verteilungen bei den Subtests Figurenzeichnen und Figurenabrufen, Wiedererkennen der Wörter und Benennen auf. Erreicht ein Proband den größtmöglichen Rohwert, wird sein Altersstandardwert deshalb als Bereich angegeben (Beim Figurenzeichnen z. B. >112). Die Messbereiche nach unten sind hingegen, wie das bei einem Test für kognitive Beeinträchtigungen sein soll, meistens zufriedenstellend.



Abbildung 5: Behandlung von Globalwerten am Beispiel der WAIS-IV

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

Wie bei vielen anderen Intelligenztestbatterien ist es auch bei den Wechsler-Tests üblich, einen Indexwert für einige (z. B. Verbalteil, Arbeitsgedächtnis) oder alle Subtests (Gesamt-IQ) zu bilden. Diese Indexwerte werden nicht als Mittelwerte der einbezogenen Subtests

ausgedrückt, sondern die Wertpunktsummen werden getrennt an der Eichstichprobe standardisiert. 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 ist das anders: Hier liegt der Fokus auf der differenzierten Erfassung von Einzelleistungen, nicht auf pauschalen Begabungskennwerten.

Die separate Standardisierung von Einzelleistungen und Globalwerten hat nun gewisse psychometrische Eigenheiten, die ausführlicher in der Dokumentation dargestellt werden. Hier soll lediglich am Beispiel der WAIS-IV erläutert werden, wie TDB2Online mit Globalwerten umgeht. **Abbildung 5** zeigt das Testprofil eines niedrig begabten Probanden, der in allen Kern-Subtests der WAIS-IV jeweils genau 4 Wertpunkte erreicht hat. Bei einem jungen Erwachsenen entspricht das einem ASW von 70 in allen Subtests. Für die Globalwerte, bei der WAIS-IV also die vier Indizes und den Gesamtwert, werden diese ASWs einfach gemittelt. Auf dem Profil erscheinen sie unter der Überschrift *Mittlere Leistungswerte* und sie haben alle den ASW 70, wie die Subtests.

Bei der Auswertung der IQs nach Handbuch (Indexwerte und Gesamttest) hätten wir andere Werte erhalten: Bei den *Sprachlichen Fähigkeiten* 67, beim *Visuellen Problemlösen* 65, im *Arbeitsgedächtnis* 66 und in der *Arbeitsgeschwindigkeit* 68. Die getrennte Standardisierung der Globalwerte ist dafür verantwortlich, dass bei ihnen eine Spreizung der Metrik erfolgt. Sie stellen nicht dar, welche Leistungen in den einbezogenen Subtests im Durchschnitt erreicht werden, sondern sie ermitteln einen Standardwert, der darstellt, wie wahrscheinlich es ist, dass jemand in **allen** einbezogenen Subtests eine (in unserem Beispiel) derart niedrige Leistung erzielt. Die Interkorrelation der Subtests untereinander und die Anzahl der einbezogenen Subtests entscheiden dann darüber, wie weit die Metrik von der der Subtests abweicht. Wir sehen, dass der Effekt bei den relativ homogenen Indizes nicht allzu auffällig ist. Der Gesamt-IQ, der mehr Subtests einbezieht, die dann auch weniger hoch miteinander korrelieren, wäre nach Handbuch im vorliegenden Fall nur noch 60, also schon deutlich niedriger als der mittlere Leistungswert. Ein so berechneter Gesamt-IQ beantwortet Fragen, die bei neuropsychologischen Untersuchungen niemand stellt.