Prof. Dr. Rolf R. Engel
University Hospital, LMU Munich, Department of Psychiatry and Psychotherapy
Nussbaumstr. 7, D-80336 Munich, Germany, ml-set@med.uni-muenchen.de

Presentation of Test Results in TDB2Online

Revision date: 13 October 2026

This document gives a concise explanation of the concepts and procedures used in TDB2Online, in particular the properties of the chosen visualization. Detailed documentation is available on the TDB2Online website, to which reference is made occasionally. The specific documentation of the individual tests is only accessible to registered professional users.

Types and Properties of Test Scores

Raw scores are the elementary values used to evaluate performance in a test. They depend on the type of task. Examples are:

- Number of correct answers in a knowledge test (Information in the WAIS)
- Number of errors in a categorization task (Halstead Category Test)
- Time in seconds needed to complete a visuoconstructional task (Trail Making Test)
- Number of correctly recalled words in a memory test (CERAD Word List)
- Number of correctly recognized words minus number of falsely "recognized" words in a recognition task (Word List Recognition in the CERAD)
- Points in a complex task scored for accuracy and speed of completion (Block Design in the WAIS)

It is immediately apparent that raw scores cannot be used to compare performance across different tests. The points in Block Design have no numerical relationship to the times in the Trail Making Test; moreover, they are scored in opposite directions.

Raw scores are, however, suitable for comparisons between individuals, and they are also commonly used for comparisons between groups (particularly when these are comparable with respect to other parameters).

Performance scores (PS) serve to make performance comparable across different tests and types of raw scores. They provide an absolute yardstick for assessing performance in a standardized form. David Wechsler was the first to apply this principle, in order to allow performance comparisons between the subtests of his intelligence batteries (something that is not possible with raw scores). Performance scores are ideally standardized on a group of young adults, roughly between 20 and 30 years of age. The reason for choosing this group is that individual cognitive performance on standardized tests generally peaks at this age (with the exception of knowledge tests, where the maximum is reached later). In younger and older people performance declines, but not to the same extent for different cognitive abilities. Other age groups are therefore unsuitable for standardization if performance is to be compared across different abilities.

A**ge-normed scores** were introduced to facilitate the interpretation of an individual patient's performance relative to his or her age cohort. They, too, are standard scores, just like performance scores. The difference is that the reference basis for the standardization comes from the part of the normative sample that is comparable in age to the patient. These scores are of little use for an absolute assessment of performance across different tests, because different abilities change differently over the lifespan. They are, however, an important aid in interpreting

an individual patient's performance relative to his or her age cohort. The clinical assessment in the written test report is based essentially on the **age-normed scores (ANS)**. Conventional manual scoring of many tests usually yields only age-normed scores in addition to the raw scores, but no age-independent standardized performance scores. Nevertheless, one should be aware that, strictly speaking, age-normed scores leave the level of ability measurement and enter the level of interpretation. More on this here.

Properties of the Scale Used

Both the performance scores and the age-normed scores use scales derived from the normal distribution. The abscissa values of the normal distribution itself (*z-scores*) are not well suited for communication (decimals, negative numbers), which is why only scales derived from them have been used for decades. TDB2Online always uses the IQ scale, with a mean of 100 and a standard deviation of 15, for both the performance scores and the age-normed scores.

The properties of the IQ scale can be seen in **Figure 1**. Most important for interpretation is the comparison between the values of the IQ scale and those of the percentile rank scale: a value of 130 on the IQ scale corresponds to a percentile rank of 97.7, which simply means that 97.7 percent of the reference sample have a score lower than 130 and the remaining 2.3 percent a higher one. For every value on the IQ scale there is a corresponding percentile rank, which can be looked up in more detailed tables.

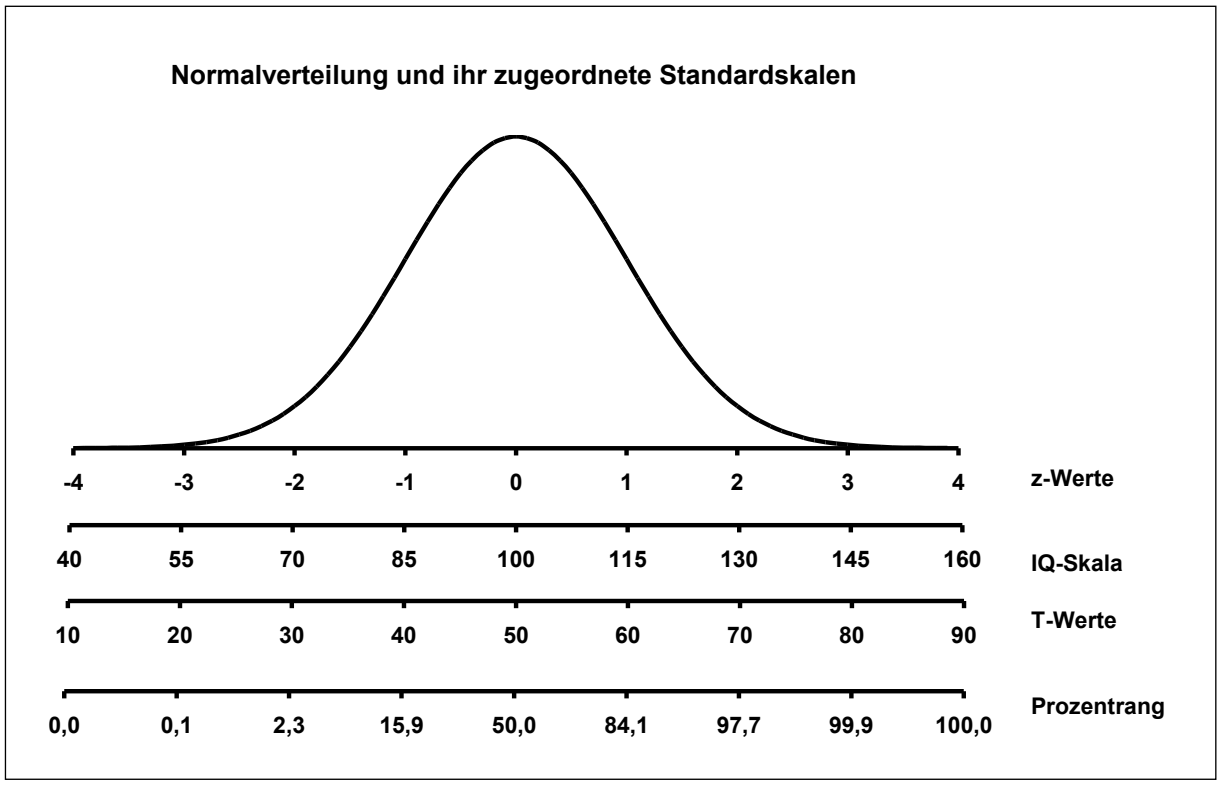


Figure 1: The normal distribution and some of the standard scales derived from it

The IQ scale has the convenient property that certain ranges of the scale lend themselves to easily remembered and interpretively meaningful verbal descriptions. About 50 percent of the normative sample lie between IQ scores of 90 and 110. In test reports this range is described as "average", sometimes with additional qualifications ("lower average range", "at the upper end of the average range", or similar). The IQ range of 80–90 comprises about 16 percent and is described as "low" or similar; below it, 70–80 is the "very low" range (about 7 %), and below that (under 70) the "extremely low" range, which statistically accounts for about 2.3 %. The upper range is divided correspondingly; **Table 1** gives an overview.

Table 1: Interpretive ranges of the IQ scale

Range	%	Qualification
below 70	2.3	extremely low
70–80	6.9	very low
80–90	16.1	low
90–110	49.5	average
110–120	16.1	high
120–130	6.9	very high
above 130	2.3	extremely high

Unfortunately, the raw scores of some tests are not normally distributed closely enough for the above assumptions and percentages to apply. In most cases the scale then has to be linearized. More on this in the detailed documentation under the heading linearity assumption.

Visualization of Test Results

The TDB2OnlineApp is a web application whose distinctive features include a uniform graphical preparation and presentation of test results. The method used for this has been in use since 2009 at the Department of Psychiatry and Psychotherapy, University Hospital, LMU Munich. It conveys information that goes far beyond a single standard score, and it has made the written description of test findings easier and more consistent. Its special features are explained below.

Reference Group for the Performance scores

As mentioned above, the ideal reference group for deriving performance scores would be a large sample of young adults aged 20–24. Unfortunately, such a group is not available for all tests. For the neuropsychological tests developed specifically for dementia assessment, such as the CERAD battery, there are no normative data for young adults. But for some other tests, too, no ideal reference group is available. The TDB2Online test profile therefore always shows, in the title of each test, the age group that serves as the reference for the performance scores. For the WAIS-IV in **Figure 2**, for example, this reads: LW-Basis: 20-29 (PS reference group aged 20–29).

Assessing Absolute Performance

In the test profile of the TDB2OnlineApp, each test score is plotted as a black dot on a performance score scale that is the same for every test and on which performance can be compared across tests. The scale with the small label "Leistungswert" (performance score) runs from 40 to 145 for every test, and the values are always in the same position. The thin vertical lines each mark one standard deviation. The numbers inside the rectangle representing a single subtest indicate the positions of the subtest's raw scores. **Figure 2** shows that a raw score of 30 on the Vocabulary Test WST corresponds approximately to a performance score of 100, and a raw score of 20 approximately to a performance score of 85. The performance scores themselves are comparable from test to test. This makes it possible to state that this 24-year-old patient's performance in recalling a word list (AVLT, labelled VLMT in the figures, *recall after interference*, raw score 12, performance score about 104) is about 20 IQ points (more than one standard deviation) better than his performance on Part B of the TMT (raw score 79, performance score about 84). Without the common measurement level provided by the performance scores, comparative interpretation of performance in different domains would not be possible.

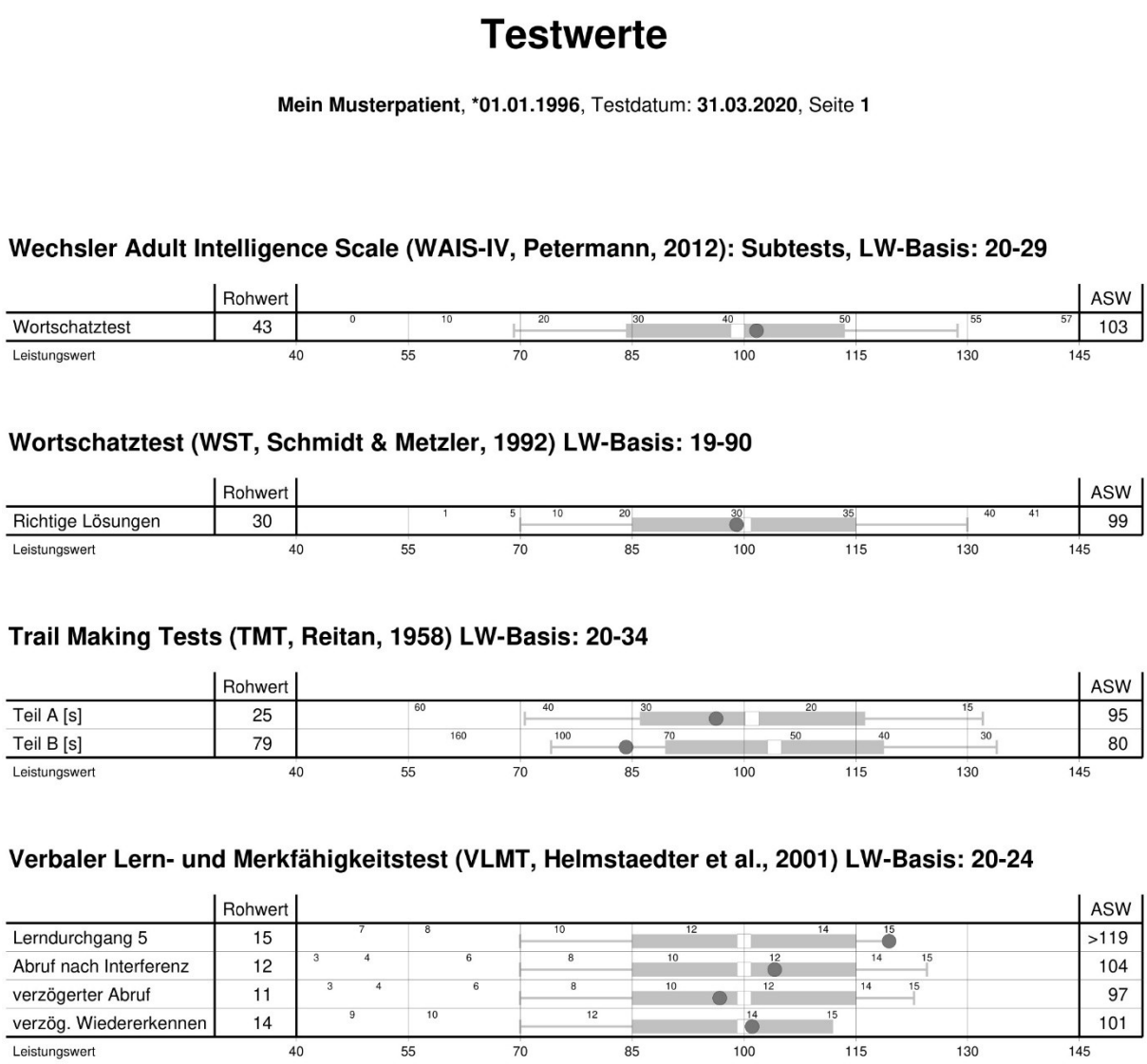


Figure 2: Sample profile of a 24-year-old patient

Reading the Measurement Range

The lowest and highest raw scores a test can yield are always marked in the profile if they fall within the displayable performance score range between 40 and 145. Using the Vocabulary Test WST in **Figure 2** as an example, one can see that the lowest possible raw score is 1 and the highest possible raw score is 41. This corresponds to a relatively wide measurement range, because the corresponding performance scores range from about 60 to about 140. The AVLT subtest *delayed recognition* (verzögertes Wiedererkennen) is different. This subtest reaches its test ceiling (the highest possible raw score of 15) at a performance score of only about 113. The potential measurement range, on the other hand, would extend very far downwards: raw scores below 9 are no longer covered by the display range of the test profile at all (and are no longer normed either, because they occur extremely rarely).

Evaluating Measurement Precision

A first estimate of measurement precision can be obtained by comparing the resolution of the raw score scale with the resolution of the performance score scale. For the Vocabulary Test WST (see **Figure 2**), raw scores between 10 and 20 correspond approximately to performance scores between 75 and 85. At this point of the scale, one additional correct word thus raises the performance score by one point, which means high measurement precision. In the upper performance range (for which the test is not primarily intended), one additional correct word

(e.g., from raw score 40 to 41) raises the performance score by 6 or 7 points. If a single additional word that is known (or even guessed) can change the performance score this much, measurement precision at this point of the scale is rather low. This example shows in very practical terms that the measurement precision of a test is by no means the same at all points of the scale (as reliability theory and the concept of the standard error of measurement derived from it suggest). Most tests measure more precisely in the middle range than at the extremes, a conclusion regularly reached with item response theory, a scaling method for test items.

One glance at the AVLT memory test (see **Figure 2**) is enough to see that all four subtests measure only relatively coarsely: a single additional word learned or recalled amounts to many IQ points on the performance score scale.

The performance score is not reported numerically; it can only be read from the position of the black dot in the test profile.

Comparison with the Individual Age Norm

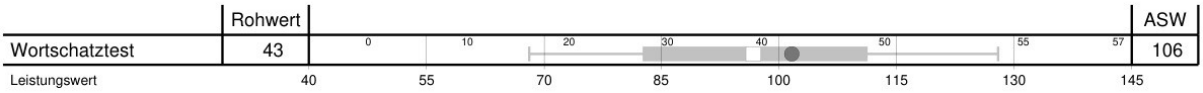
For the clinical assessment of a test score, a comparison with the performance of the corresponding age group is necessary, especially in older patients. Since the scores in the test profile are displayed graphically as age-independent performance scores, the reference ranges of the age group must be plotted in addition. The obvious choice of reference values would be the mean and the standard deviation of the age group. Unfortunately, many raw scores of psychological tests are not normally distributed (either because the distribution itself is skewed or because the measurement range ends arbitrarily at some point, truncating the distribution). If means and one- or two-sigma limits were plotted, at least the two-sigma limits would often extend far beyond the available measurement range and indicate meaningless cut-off values. In the TDB2Online profile, reference ranges determined via percentile ranks (PR) are therefore plotted. The white square marks the median (= PR 50) of the patient's age group, the grey rectangle the range from PR 16 to PR 84, and the two whiskers at the ends of the grey rectangle percentile ranks 2.5 and 97.5. If the performance scores are normally distributed, which can often be achieved by linearizing the raw scores, these percentile ranks correspond to the mean and the one- and two-sigma limits.

A patient's individual position within the limits of his or her age reference group can be read from the graph by looking at where the black dot lies in relation to the grey rectangle and the outer whiskers. This would, however, be somewhat laborious. The program therefore calculates the age-normed scores and displays them on the right-hand side of the profile in the column *ASW* (= ANS, age-normed score). These age-normed scores are also displayed on the IQ scale, i.e., with a mean of 100 and a standard deviation of 15. An example shows how the score is calculated: in the first row of the Trail Making Test in **Figure 2** (*Teil A* = Part A), one can see that the patient's individual test score lies about one third of the grey area to the left of the white square. If it were on the white square, this would correspond to an age-normed score of 100. If it were at the left end of the grey area, this would correspond to an age-normed score of 85. Since it lies about 1/3 of the grey area below the white square, this corresponds to an age-normed score of 95. This score is entered in the *ASW* column on the right.

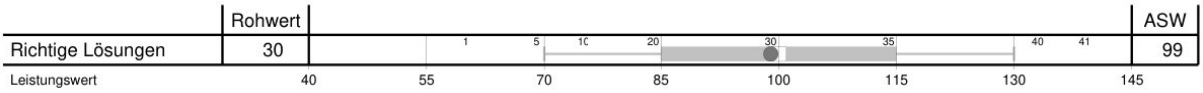
Testwerte

Mein alter Musterpatient, ***01.01.1946**, Testdatum: **31.03.2020**, Seite **1**

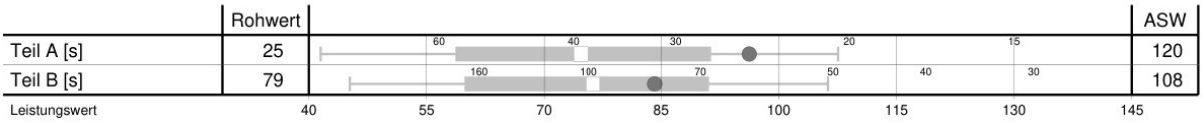
Wechsler Adult Intelligence Scale (WAIS-IV, Petermann, 2012): Subtests, LW-Basis: 20-29



Wortschatztest (WST, Schmidt & Metzler, 1992) LW-Basis: 19-90



Trail Making Tests (TMT, Reitan, 1958) LW-Basis: 20-34



Verbaler Lern- und Merkfähigkeitstest (VLMT, Helmstaedter et al., 2001) LW-Basis: 20-24

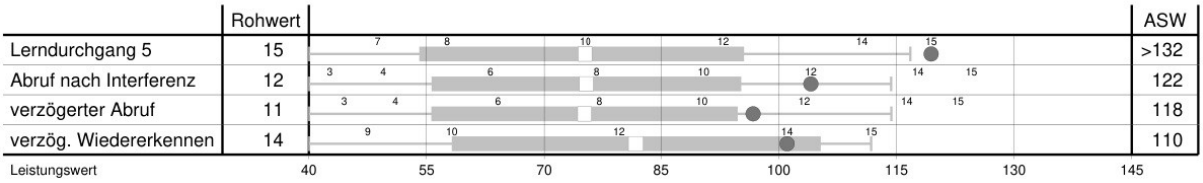


Figure 3: Test performance as in Figure 2 for a 74-year-old patient

In young adults, the age-normed scores hardly differ from the performance scores. With normally distributed raw scores, the grey area generally ranges from 85 to 115, and the whiskers extend down to 70 and up to 130. The older a patient is, and the more performance on a particular test declines with increasing age, the greater the difference between performance scores and age-normed scores. This can be seen in **Figure 3**, which shows the same test scores as **Figure 2**, but this time for a patient aged 74 instead of 24. The position of the black dots in the performance score grid remains the same. What changes are the age reference ranges plotted in the profile and, accordingly, the age-normed scores in the ASW column. The patient's different age has a very strong effect on the visuoconstructional Trail Making Test (TMT; here the age-normed score of **95** for Part A in Fig. 2 becomes an ASW of **120** in Fig. 3) and on the AVLT memory test. It has no effect on the Vocabulary Test WST, because there are no age-graded norms for this test and the age of 74 is still within the age range of the normative sample. In the Vocabulary subtest of the WAIS-IV, one can see in a little more detail that at the age of 74 there are also slight age-related declines in vocabulary in the normative sample, which give our examinee a somewhat higher ASW of 106 for the same performance. It is quickly apparent, however, that the age effects on the TMT and the AVLT are much larger.

Limits of the Measurement Range

Not all tests cover a wide or even a symmetrical performance range. The measurement range of the WAIS-IV subtest *Information* is exemplary, extending from about performance score 47 to 143 (see **Figure 2 or 3**). The AVLT subtests, by contrast, cannot measure very high performance. For all subtests, 15 is the highest attainable raw score; depending on the subtest, it corresponds to performance scores between 112 and 124. For the first subtest, learning the words, shown in the profile in **Figures 2 and 3** as *learning trial 5* (Lerndurchgang 5), a raw score of 15 leads to a performance score of about 120. For a 24-year-old (**Figure 2**), this also corresponds to the age-normed score; for a 74-year-old (**Figure 3**), by contrast, the age-normed score is about 133. In both cases, however, a raw score of 15 marks the end of the measurement range. To show that higher performance cannot be measured, the age-normed scores are given in such cases as ranges open at the top (>119 and >132, respectively, in the example). The same procedure is applied when the measurement range of a test is limited at the lower end.

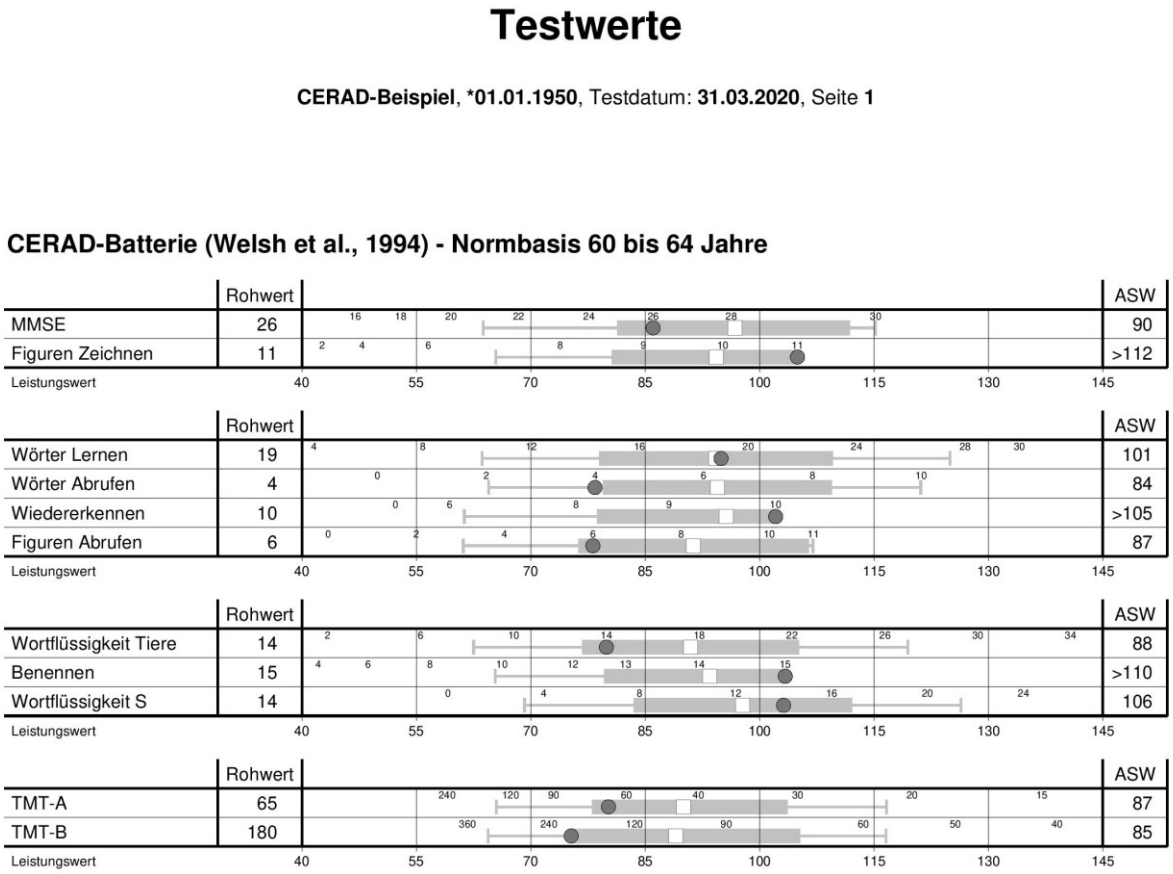


Figure 4: CERAD test scores of a 70-year-old

Special Features of Tests for Dementia Assessment

As mentioned above, many tests for dementia assessment have no norms for young adults. A prominent example of such tests is the CERAD test battery. **Figure 4** shows CERAD test results for a seventy-year-old patient. Here, too – similar to the AVLT in Figures 2 and 3 – the restricted measurement ranges and the resulting extremely skewed distributions are conspicuous in the subtests Constructional Praxis and Constructional Praxis Recall (Figuren Zeichnen, Figuren Abrufen), Word List Recognition and Naming. If an examinee achieves the highest possible raw score, his or her age-normed score is therefore given as a range (for Constructional Praxis, e.g., >112). The measurement ranges at the lower end, on the other hand, are mostly satisfactory, as they should be in a test for cognitive impairment.

Testwerte

Wenig Begabt, ***01.01.1997**, Testdatum: **26.03.2020**, Seite **1**

Wechsler Adult Intelligence Scale (WAIS-IV, Petermann, 2012): Subtestleistungen

	Rohwert	ASW
Gemeinsamkeitenfinden	9	70
Wortschatztest	18	71
Allgemeines Wissen	4	70
Allgemeines Verständnis	12	70
Leistungswert	40 55 70 85 100 115 130 145	

	Rohwert	ASW
Mosaiktest	25	70
Matrizentest	11	69
Visuelle Puzzles	8	70
Formenwaage	8	70
Bilderergänzen	7	71
Leistungswert	40 55 70 85 100 115 130 145	

	Rohwert	ASW
Zahlennachsprechen	17	69
Rechnerisches Denken	8	71
Buchst.-Zahlen-Folgen	13	70
Leistungswert	40 55 70 85 100 115 130 145	

	Rohwert	ASW
Symbolsuche	18	70
Zahlen-Symbol-Test	37	70
Durchstreichtest	24	70
Leistungswert	40 55 70 85 100 115 130 145	

WAIS-IV: Mittlere Leistungswerte

	Rohwert	ASW
Sprachl. Fähigkeiten	70	70
Visuelles Problemlösen	70	70
Arbeitsgedächtnis	70	70
Arbeitsgeschwindigkeit	70	70
Leistungswert	40 55 70 85 100 115 130 145	

	Rohwert	ASW
Gesamttest	70	70
Leistungswert	40 55 70 85 100 115 130 145	

Figure 5: Handling of global scores, using the WAIS-IV as an example

Procedure for Global Scores (e.g., Full Scale IQ)

As with many other intelligence test batteries, it is customary with the Wechsler tests to form an index score for some subtests (e.g., verbal comprehension, working memory) or for all of them (Full Scale IQ). These index scores are not expressed as means of the included subtests; instead, the sums of scaled scores are standardized separately on the standardization sample. This practice goes back to the beginnings of intelligence testing, when the main aim was to assess global intellectual ability rather than its structure. In current neuropsychological assessment this is different: the focus is on a differentiated assessment of individual abilities, not on global ability indices.

The separate standardization of individual abilities and global scores has certain psychometric peculiarities, which are described in more detail in the documentation. Here we merely explain, using the WAIS-IV as an example, how TDB2Online handles global scores. **Figure 5** shows the test profile of a low-ability examinee who obtained exactly 4 scaled score points on every core subtest of the WAIS-IV. For a young adult, this corresponds to an age-normed score (ASW) of 70 on all subtests. For the global scores – in the WAIS-IV the four indices and the full scale score – these ASWs are simply averaged. In the profile they appear under the heading *Mittlere Leistungswerte* (mean performance scores), and like the subtests they all have an ASW of 70.

Scoring the IQs according to the manual (index scores and full scale) would have produced different values: *Verbal Comprehension* 67, *Perceptual Reasoning* 65, *Working Memory* 66 and *Processing Speed* 68. The separate standardization of the global scores is responsible for a spreading of their metric. They do not show what performance is achieved on average in the included subtests; instead, they determine a standard score that expresses how likely it is that someone achieves such a low performance (in our example) on **all** of the included subtests. The intercorrelation of the subtests and the number of subtests included then determine how far the metric deviates from that of the subtests. We see that the effect is not very pronounced for the relatively homogeneous indices. The Full-Scale IQ, which includes more subtests that are also less highly correlated with one another, would in this case be only 60 according to the manual, i.e., already clearly lower than the mean performance score. A Full-Scale IQ calculated in this way answers questions that nobody asks in neuropsychological assessments.