



How to read, write and challenge a psychological test result in expert evidence

A practical guide in ten questions

Dr Milos Kankaras | Center for Psychology | October 2026

Who this is for. Judges, prosecutors and lawyers who receive an expert opinion that rests on a psychological test, and psychologists who write such opinions. It is free and may be shared in full.

What it is not. It is not legal advice, not an expert opinion in any case, and it describes no real case. All numerical examples are invented and serve only to show the arithmetic.

What this guide is about

In expert evidence a psychological test usually arrives as one number: a score of 68, the 93rd percentile, "below the cut-off". The number looks precise. Behind it sits a chain of decisions that a court rarely sees: which test was chosen, in which language it was given, which group it was compared with, where the cut-off was drawn, and how different the score might have been had the same person been tested again on the same day. This guide describes that chain and puts ten questions that test it.

The law of expert evidence differs between jurisdictions, and so do the rules on disclosing test materials, keeping records and the limits of an expert's opinion. The guide does not describe any of them. It deals with the measurement side of a test-based opinion, which is the same wherever the opinion is given, and it names the points where local rules will decide what is permitted.

How to use the guide

Each of the ten questions has the same parts: why it matters, what to ask, what a good answer looks like and what the warning signs are. A one-page checklist at the end can be carried to a hearing or kept with the opinion. The guide reads in order, but every chapter also stands alone.

Three terms are used consistently. An *instrument* is a test, questionnaire or scale. A *score* is the number the instrument gives. A *conclusion* is the claim the expert draws from the score. Most disputes arise not over the score but over the path from score to conclusion.

What the guide does not do

It does not assess any expert, recommend any test, or say what is legally relevant in a given proceeding. That judgement belongs to the court. It does not replace professional supervision or the opinion of a second expert.

What a test score is

Three properties of a score need to be clear before the ten questions, because everything else is built on them.

A score is a sample of behaviour, not insight into a person. A test observes responses in one situation, at one moment, on a particular set of tasks. From these it infers a property that cannot be seen directly, such as reasoning ability, the level of anxiety or a personality trait. Every such step is an inference, not an observation, and it needs evidence that it is justified.

A score contains measurement error. The same person, tested twice under the same conditions, will not get the same number. The difference is not an expert's mistake but a natural property of measurement, and it can be estimated. A score without a range is therefore an incomplete piece of information.

A score is interpreted by comparison. On its own the number 52 means nothing. It acquires meaning only by comparison with norms (how similar people perform), with a cut-off (where scores are sorted into categories) or with an earlier score of the same person. Each comparison rests on assumptions, and the assumptions are usually where the dispute lies.

The Standards for Educational and Psychological Testing, issued jointly by three American associations for educational research, psychology and measurement, start from the same view: validity is not a property of a test but of the interpretation of scores for a specific purpose and population (American Educational Research Association et al., 2014). The ten questions that follow develop that sentence.

Question 1: Is the instrument valid in this language, for this purpose and this population?

Why it matters

An instrument that was carefully developed and checked in one country and language does not become valid by being translated. Translation changes the meaning of individual statements, and culture changes how people respond to them. The International Test Commission therefore separates simple translation from adaptation, and asks for evidence of validity to be collected for each new language and cultural setting rather than assumed to carry over from the original (International Test Commission, 2018).

The second limit is purpose. A questionnaire built to monitor mood in a clinic has not been checked as evidence of someone's parenting ability, and a test of working memory has not been checked as evidence of fitness for work. The question is not whether the test is good, but whether it is good for the question the court has asked.

What to ask

- Exactly which version of the instrument was used (edition, year, language), and who made the translation?
- How did the translation come about: translation and back-translation, expert review, pilot administration? Did the rights holder approve the adaptation?
- Is there published evidence of reliability and validity for this language and country? Where is it published?
- Does the evidence apply to the population the person belongs to (age, education, language spoken at home)?
- Was the instrument checked for the purpose it was used for here, or only for a related one?

What a good answer looks like

A good opinion names the instrument and version, cites a source showing reliability and validity in this language and population, and says openly if no such source exists. Where there is none, honesty is the standard: the score is reported as orientation, and the conclusion is limited to what can be supported without the instrument by other sources. Whether an instrument works equally well in different groups is checked by testing measurement invariance, that is, whether the same concept is measured in the same way in every group; a review of the literature shows that the conventions for carrying out and reporting these tests are still developing (Putnick & Bornstein, 2016). The expert need not compute this personally but should know whether such a test has been published.

Warning signs

- The statement "the test is valid" with no source, or a source that concerns another language, another country or another purpose.

- A translation with no record of how it was made, or one that the expert made personally "for the needs of the case".
- An instrument developed for screening or clinical use presented as evidence on a legal or functional question.
- A scale whose name matches a concept in the dispute while its item content was never checked. A scale called "parenting capacity" does not necessarily measure what the name promises.

Question 2: Which norms were used, how old are they and on whom were they collected?

Why it matters

A percentile, a T-score and a standard score are not properties of a person but a relation between a person and a group. The same raw score gives different percentiles when compared with a student sample, with the general adult population or with a clinical group. Choosing the norm means choosing the outcome of the comparison.

Example A (invented). A person has a raw score of 52 on the scale ANX-20. Norm 1, obtained on students: mean 40, standard deviation 8. Norm 2, obtained on adults from the general population: mean 45, standard deviation 9.

Against norm 1, $z = (52 - 40) / 8 = 1.50$, about the 93rd percentile. Against norm 2, $z = (52 - 45) / 9 = 0.78$, about the 78th percentile.

The same number gives two conclusions: "among the most pronounced seven percent" and "somewhat above most". An opinion that does not say which norm was used can be neither checked nor challenged.

The age of the norm is a separate question for ability tests. Average scores on intelligence tests rose from generation to generation during the twentieth century. A meta-analysis covering studies from 1909 to 2013 found an average gain of about three IQ points per decade, with differences between countries and types of test and signs of slowing in more recent times (Pietschnig & Voracek, 2015). The practical consequence is that a test with an old norm can systematically overstate the performance of someone tested today. The size of the effect depends on the test and must not be assumed, but it has to be asked about.

What to ask

- Which norm was used: the title of the manual, edition, year, table?
- On whom was it collected: country, age, sex, education, language, sample size, the way participants were selected (random sample or convenience sample)?
- Do norms exist for the country the person lives in? If not, why was this foreign norm chosen and how similar are the populations?
- Is the person comparable with the norm group in age and education, and if not, how was that taken into account?
- How old is the norm?

What a good answer looks like

The opinion cites the norm so that another expert can find it: manual, year, table, description of the sample. Where no local norm exists, it says so, explains why a foreign norm is acceptable and notes that scores should be read as approximate. A good opinion does not report percentiles with more precision than the norm allows.

Warning signs

- A percentile or T-score with no norm named.
- A norm from a manual whose sample is not described, or a sample that clearly differs from the person (for example students for someone who left school at sixteen).
- A norm from a standardisation several decades old, with no discussion of its age.
- A foreign mean and standard deviation applied to a translated instrument as if it were the same measurement.
- A norm for one age band applied to a person at the edge of the band, without noting that this decided the outcome.

Question 3: What exactly does the cut-off mean and where does it come from?

Why it matters

A cut-off turns a continuous score into a category: "above" or "below", "clinically significant" or "not". The category is an agreement, not a fact about nature. Where the line is drawn depends on purpose: for a first screen it is set low to catch as many people as possible, and for a decision with heavy consequences it is set high to avoid wrong claims. A cut-off chosen for one purpose does not transfer to another.

The second matter is what a cut-off says about one person. How well it works is described by sensitivity (the share of people with the property that it finds) and specificity (the share of people without the property that it correctly leaves out). But the probability that a person above the cut-off really has the property also depends on how common the property is in the group being tested. This is a classic finding, known since the middle of the last century, about how the frequency of a property in the population affects the usefulness of cut-offs (Meehl & Rosen, 1955).

Example B (invented). A scale has sensitivity 0.85 and specificity 0.80. 1,000 people are tested.

If the property is present in 5% of them (50 people), the scale finds 42.5 of them, and among the 950 without the property it wrongly flags 190. About 232 people are flagged and 42 of them have the property. The probability that a person above the cut-off really has the property is about 18%.

If the property is present in 30% (300 people), the scale finds 255, and among the 700 without it wrongly flags 140. Of 395 flagged, 255 have the property, about 65%.

The same scale, the same cut-off, the same sensitivity and specificity, yet "above the cut-off" means something quite different in the two groups. The question "how common is the property in the group the person comes from" is statistical, not legal, and the opinion should answer it.

What to ask

- Who set the cut-off and on what basis: a manual, a study, custom?
- For what purpose was it derived (screening, diagnostic confirmation, research)?
- What are its sensitivity and specificity, and in which population were they measured?
- What is the estimate of how common the property is in the group the person comes from, and how was it taken into account?
- Where does the score lie relative to the cut-off, and does the confidence interval (Question 4) leave it on one side of the line?

What a good answer looks like

The opinion says that the cut-off is a convention, gives its origin and purpose, claims no more than sensitivity and specificity allow and, when the score is close to the line, says expressly that the classification is not reliable. The statistical fact "the score is below the cut-off" stays separate from the

legal question of what that fact means for the outcome of the proceeding. The second belongs to the court (Question 9).

Warning signs

- A cut-off introduced with no source, or one from a research study and used as diagnostic.
- A score of 68 presented as "below 70" without considering that measurement error does not reliably separate it from 70.
- Two categories ("within normal limits" and "pathological") presented as two kinds of people rather than two parts of one scale.
- Sensitivity and specificity taken from a study of a clinical group and applied to the general population (or the reverse).

Question 4: What is the standard error of measurement, and is the score reported with a range?

Why it matters

Reliability says how consistent scores are on repetition. The standard error of measurement (SEM) turns that coefficient into the units of the test itself: how far a score is expected to move because of chance factors such as the day's form, attention or luck in choosing answers. It is computed from the standard deviation of the scale (SD) and the reliability coefficient (r): $SEM = SD \times \text{the square root of } (1 - r)$. The interval within which a score can be expected about 95% of the time is approximately the score plus or minus 1.96 SEM. This is a simple approximation, and the more elaborate procedures that the Standards describe are sometimes more appropriate (American Educational Research Association et al., 2014), but it shows the scale of the problem well enough.

Example C (invented). A test has mean 100 and SD 15. Its reliability is $r = 0.95$. $SEM = 15 \times \text{the square root of } 0.05$, which is about 3.4 points.

A person scores 68. The 95% interval is 68 plus or minus 6.6, that is, from about 61 to about 75. The cut-off of 70 lies inside that range.

The statement "the score is below 70" is true of the measured number, but it does not reliably separate this person from people whose true level is above 70. Had the reliability been 0.90, the SEM would be about 4.7 and the interval about plus or minus 9.3 points.

Two consequences deserve attention. First, subscale scores and differences between two scores carry more error than a total score, and should not be discussed with the same certainty. Second, the reliability coefficient in a manual applies to the group on which it was computed. If the person differs from that group (in language, age or condition), the coefficient may be different.

What to ask

- Which reliability coefficient was used, of what type (internal consistency, test-retest, raters) and on which group was it obtained?

- What is the SEM for this score and in which units?
- Is the score reported with an interval rather than only as a point value?
- If scores are sorted into categories, does the interval cross the boundary of the category?
- For subscales or differences between scores: is their reliability stated separately?

What a good answer looks like

Every score on which a conclusion rests is stated with an interval, and a classification near a boundary is described as conditional. Reliability is taken from a source, not from memory, and the source is named.

Warning signs

- Scores reported to two decimal places, as though the accuracy were greater than the test permits.
- A category ("average", "below average") with no range at all.
- A conclusion drawn from the difference between two subscores, such as "verbal better than non-verbal", without checking whether the difference exceeds the error.
- A reliability coefficient with no source, or from a group that clearly differs from the person.

Question 5: Has the person been tested before with this or a similar instrument, and what does that mean for the score?

Why it matters

Repeated testing with the same instrument usually produces a better score, because the person remembers the tasks, knows what is expected or has learned a strategy. This is called a practice effect. Meta-analyses of neuropsychological tests show that its size depends on the interval between testings, on the age of the participants and on whether an alternate form of the test is used (Calamia et al., 2012). In disputes this is not a theoretical point: the person may have been tested with the same instrument by an earlier expert, by a physician or in an earlier proceeding, and the question "how many times and when" is often never asked.

The same question works in the other direction. When a score is compared with an earlier one to conclude that something has changed, one needs to know whether the difference is larger than measurement error alone would produce. The standard measure is the reliable change index, which divides the difference between two scores by the standard error of that difference (Jacobson & Truax, 1991).

Example D (invented). The same test as in Example C (SEM 3.4). First testing: 68. Second, five weeks later: 74. The difference is 6 points.

The standard error of the difference = the square root of $2 \times \text{SEM}$, about 4.7. The index = $6 / 4.7$, about 1.27. A value below 1.96 means that the difference cannot be reliably distinguished from measurement error alone, even before allowing for the possibility that part of the rise was a practice effect.

The claim "the condition has improved" is not supported by these numbers, and the claim "the condition has not changed" is not proved either: all that has been shown is that this test cannot reliably establish a change.

What to ask

- Has the person solved this or a similar test before? When, where, for what purpose, and are the earlier scores available?
- What is the interval between testings, and is there a parallel form?
- Is the change in score larger than measurement error explains?
- How was regression to the mean taken into account (extreme values tend to move towards the average on retesting)?
- What changed in the meantime in the person and in the circumstances (treatment, sleep, medication, motivation)?

What a good answer looks like

The opinion says whether earlier testings are known and, if not, that they could not be excluded. Changes are interpreted with an interval, not as a simple difference. A comparison of scores from different versions or different examiners carries a note that they were not equated.

Warning signs

- The opinion does not mention earlier use of the same instrument although the person has been in proceedings for a long time.
- The difference between two scores presented as "improvement" or "deterioration" without reliability.
- Scores from different editions of the same test compared as though they were one measurement.
- A conclusion about a lasting condition drawn from one testing at the moment of greatest pressure.

Question 6: How were response style and possible malingering checked?

Why it matters

In a clinical setting the person usually wants to be described accurately. In a legal setting the outcome of the proceeding may depend on the result, so there are incentives to exaggerate, to play down or to misrepresent deliberately. The test itself does not know why the person answers as they do. This is why indicators are sought, alongside the score, that responding was consistent and that real effort was made.

There are two kinds of check, and both have limits. Validity scales built into questionnaires (inconsistency, exaggeration, socially desirable responding) show the manner of responding. Separate effort and performance validity tests check whether the person did the best they could on ability tasks. Both approaches are an integral part of serious assessment in neuropsychology (Larrabee, 2012), and a comprehensive review of the procedures and their limits is found in the handbook on the assessment of malingering and deception (Rogers & Bender, 2018).

What the checks cannot do matters as much. Failing an effort test does not prove intent, because the same pattern can be produced by pain, fatigue, misunderstanding of the instructions or weak language skills. A normal result on a validity scale does not prove that responding was honest. No test uncovers a lie about an event, only the manner of working on the test. The conclusion "malingers" or "does not malingering" is therefore a conclusion about intent, which a psychological score cannot carry on its own.

What to ask

- Which checks of response style or effort were carried out and what were their results?
- If none were, why not, and how did that limit the conclusion?
- What other than intent could explain the pattern of responses (language, literacy, fatigue, medication, pain, misunderstanding)?
- Are the results consistent with other sources (records, information in the file, observed behaviour)?
- What were the conditions of testing: duration, breaks, presence of other people, interpreter?

What a good answer looks like

The opinion lists every check that was carried out and its outcome, does not skip the unfavourable ones, and distinguishes a "deviating response pattern" from "deliberate deception". Where a check was not carried out, the interpretation is expressly limited. A conclusion about intent, if one is drawn at all, rests on several mutually consistent sources and not on one score.

Warning signs

- No check at all, and a conclusion stated as though responding were honest and complete (or the reverse).
- A single score below the cut-off on an effort test presented as proof of malingering.

- No discussion of language, education and conditions of testing when the person does not speak the instrument's language as a mother tongue.
- A claim that the test "uncovered a lie" or "established the truthfulness of a statement".

Question 7: Was the instrument legitimately used, and is the expert authorised to use it?

Why it matters

Published psychological tests are protected by copyright, and their publishers usually restrict sales to people with appropriate qualifications and keep the tasks and keys confidential. This is not a formality. Norms and validity evidence belong to the exact published version. An unauthorised copy, a test reproduced from memory, a translation made without the rights holder's consent or a scoring key from an unknown source may contain an error that nobody can check. The Standards expressly connect test security with the validity of score interpretation (American Educational Research Association et al., 2014).

The second matter is access by the court and the parties. A court is entitled to check what an opinion rests on, while the tasks and keys are sensitive material whose public availability would make the test useless for every later person. The solution sought in practice is that raw material is not given to the parties freely, but is disclosed to another qualified person who is also entitled to work with that test, for example a second expert appointed by the court. The conditions vary from publisher to publisher and from jurisdiction to jurisdiction, and should be read rather than assumed.

What to ask

- Whose licence is it, and what is the proof of purchase: an invoice, an order, a licence for scoring software?
- What level of qualification does the publisher require for this test, and does the expert meet it?
- If the instrument was translated or adapted, did the rights holder authorise it?
- How were scoring and norms applied: the publisher's key, licensed software or a home-made table? Who checked the table?
- How is the court's access to raw data arranged without compromising test security?

What a good answer looks like

The expert can show, on request, where the instrument came from, which version was used for scoring and that this was permitted. Material that must not be shared publicly is available to the court in a way the publisher's terms allow. Where necessary, the opinion states which parts were left out of the public section and why.

Warning signs

- The expert cannot say where the instrument or the key came from.
- An adapted or "shortened" test used with the norms of the full test.
- Scoring in an improvised spreadsheet with no check and no documented version.
- Refusal of any access to raw data on the ground of "test confidentiality", with no way offered for a qualified person to check it.

Question 8: What can the score not show?

Why it matters

The biggest errors in the use of tests are seldom arithmetic. More often a score is credited with more than it can carry. An expert who knows the limits of the instrument does not weaken the opinion by saying so but makes it checkable. An opinion should therefore state what the score does not show.

A test score can	A test score cannot
Describe how the person worked on tasks under the conditions of testing	Prove what the person's state was at the time of the events in dispute
Show whether performance differs from that of a comparable group	Establish the cause of that difference (illness, injury, fatigue, education, language)
Be one source for a conclusion about functioning	Replace clinical observation, records and other sources
Support a description of strengths and weaknesses	Predict an individual's future behaviour with certainty
Point to a response pattern that deviates	Prove intent, lying or the truthfulness of a statement
Show how pronounced a trait is on a scale	Establish a legal category (capacity, responsibility, suitability)

Three misunderstandings recur.

From group to individual. The statement that a pattern is common among people with a given diagnosis does not mean that a person with that pattern has the diagnosis. The probability depends on how common the diagnosis is (Example B). Conversely, the fact that a group on average scores lower says nothing about one of its members.

From score to functioning. A test measures performance on a standardised task in a controlled situation. How the person copes at home, at work or with a child, a test assesses indirectly and never measures directly. The closer the question in dispute is to everyday functioning, the more sources are needed besides the test.

From the present to the past. A score is a snapshot of the moment of testing. A conclusion about the state at an earlier moment needs additional information and always remains conditional.

What to ask

- Which exact question in the instruction does the score answer, and which does it not?
- Which other sources were used, and do they agree with the test?
- How was an explanation of the pattern that does not involve the disorder or ability in dispute taken into account?
- Does the conclusion relate to the time of testing, or is it attributed to an earlier period, and on what basis?

What a good answer looks like

The opinion has an express section on limitations, stating the limits of the instrument and the limits arising from the conditions of this testing. The conclusion is confined to what the score and the other sources carry together.

Warning signs

- A conclusion about cause drawn from a score that measures only outcome.
- A claim about a past state drawn from a present score with no further sources.
- One instrument as the only support for a question about everyday life.
- No discussion at all of alternative explanations.

Question 9: Where do facts end and legal conclusions begin?

Why it matters

In many legal systems the task of an expert is to establish facts and to give a reasoned professional view on them, while the application of legal standards is for the court. Local rules differ, and the point should be checked locally. For a test-based opinion the practical meaning is this: the expert can say what was measured, how it compares with a group, how reliable it is and what follows professionally about functioning. The expert should not say whether the conditions of a legal category are met. Categories such as criminal responsibility, legal capacity or the best interests of a child are not psychological concepts, and applying them is the task of the court. A psychologist can describe the functional abilities that the court then relates to the legal standard.

The line is not always sharp, so it helps to know the patterns. A formulation that stays on the side of facts describes measurement and functioning. A formulation that crosses the line uses the word of a legal concept or announces an outcome that belongs to the court.

Formulation that stays with facts	Formulation that crosses into a legal question
"On the reasoning test the person obtained a score that, within the limits stated in the opinion, lies in the lower part of the distribution of a comparable group."	"The person is incapable of making decisions independently."
"The measured patterns of attention and memory indicate difficulty in following complex instructions."	"The person was not responsible at the time of the act."
"The results show no difficulty on tasks measuring comprehension and planning, but cannot be attributed to the time of the events."	"The person was capable of sound judgement overall."
"In assessing parental functioning the test shows the following response patterns; the test does not measure parenting ability as such."	"The parent is not fit to bring up the child."

The line works in both directions. A court can put questions to an expert so that they stay on the side of facts: "Which abilities were measured and with what reliability?" instead of "Is the person capable?" The parties, for their part, can object to an opinion that has crossed the line, because a conclusion that belongs to the court is one an expert cannot reach.

What to ask

- Does the question in the instruction call for a factual or a legal answer?
- Does the opinion use concepts that are legal rather than psychological and, if it does, are they cited only as the question the court posed?
- Is every conclusion tied to a measurement or an observation?
- Does the expert separate what was established from what is assessed as probable?

What a good answer looks like

The opinion answers factual questions in factual language, ties every conclusion to a measurement or an observation, and leaves legal qualification to the court. Where the court asks for an "opinion on capacity", the expert answers with a description of functional abilities and explains what follows from them professionally.

Warning signs

- The closing sentence uses a legal concept as its predicate.
- A conclusion with no visible support in a score or an observation.
- An opinion that answers a question the court did not ask, or fails to answer the one it did.
- The language of certainty ("undoubtedly", "obviously") where measurement permits only probability.

Question 10: What must the file contain, and can the work be checked?

Why it matters

An opinion that cannot be checked turns into a question of trust. Rules on record-keeping and retention differ between jurisdictions and professional codes: many require the opinion and the underlying records to be kept for a set period, and some prescribe a form for recording each expert engagement. Check the rules that apply to you. Whatever they say, the measurement logic is the same: a second qualified person should be able to repeat the scoring and see whether the conclusions follow from the scores.

A sensible practice is to keep the raw material with the opinion for as long as the opinion is kept, because the opinion cannot be checked without it. That is a recommendation of this guide, not a legal requirement, and it needs to be reconciled with the rules of the jurisdiction. Raw responses are personal data, often sensitive, and data protection rules apply to them.

What an opinion based on a test should contain

- **Instrument:** full name, edition, language, place and year of publication, source of the translation.
- **Circumstances:** date, start time and duration, place, who was present, the interpreter (if used), the language used, deviations from standard procedure.
- **Person:** how identified, what information they received and agreed to, their condition at the time of testing (fatigue, medication, health problems that may have had an effect).
- **Raw data:** the original protocol and answer sheets, or the technical record when the test is electronic.
- **Scoring:** who scored, with which key or software and which version, and who checked.
- **Norms:** name, year and table, a description of the sample, the reasons for the choice (Question 2).
- **Results:** raw and derived, with confidence intervals (Question 4).
- **Checks of response validity:** which were carried out and what they showed (Question 6).
- **The link to the conclusion:** which score carries which claim, and which other sources the claim used.
- **Limitations:** what the score does not show (Question 8) and what could not be checked.

What to ask

- Can another qualified person repeat the scoring from the documentation supplied and arrive at the same score?
- Is the norm used cited with page and table?
- Is the raw material kept, and how is it protected?
- Was the record that the rules require kept for this engagement?

What a good answer looks like

The opinion is such that another qualified psychologist, with access to the raw material, can reproduce the scoring and check whether the conclusions follow from the scores. If the raw material is not in the file for reasons of test security, the opinion states where it is kept and how it can be passed to a qualified person.

Warning signs

- The file holds only a table of final scores.
- Scoring was done in a file with no stated version, and the raw responses cannot be obtained.
- The date, duration and conditions of testing were not recorded.
- The opinion does not say who did the scoring.

A worked example: an invented opinion read through the ten questions

Everything in the example is invented: the person, the test, the numbers and the events do not exist and refer to no real case. It shows how the ten questions apply to one extract from an opinion.

Extract from an invented opinion. Person A. B. (aged 41) was tested on 14 March with the Cognitive Ability Scale CAS-4, in a translation prepared by the expert. The total score is 68, which is below the cut-off of 70 and indicates significantly reduced cognitive ability. The score was interpreted against the norms in the manual. The person had been tested earlier by a physician, but that score is not available. The person was cooperative, so the results are valid. On this basis, the person is not capable of understanding the consequences of signing a contract.

Sentence or decision in the opinion	Question	What to ask for
"CAS-4, in a translation prepared by the expert"	1 and 7	The source of evidence of validity in this language; whether the rights holder approved the translation; how the translation was checked
"Norms in the manual"	2	Which edition and year, on whom collected, whether they fit A. B. (age, education, language)
"68, below the cut-off of 70"	3 and 4	The origin of the cut-off; the SEM and the interval. With the reliability from Example C the interval runs from about 61 to about 75 and crosses the cut-off
"Tested earlier by a physician, score not available"	5	Which test, when, and whether it could have influenced today's score
"Was cooperative, so the results are valid"	6	Cooperation is not a validity check; which effort or consistency checks were done?
"Indicates significantly reduced cognitive ability"	8	What was used besides the test; whether there are other explanations (language, fatigue, misunderstanding the instructions)
"Is not capable of understanding the consequences of signing a contract"	9	A legal conclusion; the expert describes functional abilities and the court gives the legal assessment
The file holds only the score	10	Raw protocol, key, version, date and duration, who scored

None of these remarks claims that the conclusion is wrong. They claim only that the material supplied does not let anyone see that it is right. The questions exist for exactly that purpose: to separate "not shown" from "not true".

The same passage, written so that it can withstand checking, might look like this.

The same passage, corrected. Person A. B. (aged 41) was tested on 14 March for 55 minutes, with the

examiner present and nobody else, in the language the person speaks at home. The Cognitive Ability Scale CAS-4 was used (2012 edition, licensed copy, source attached), in a translation approved by the rights holder and with published reliability data ($r = 0.95$). The raw score corresponds to a standard score of 68 against the norm in the manual (general adult population, page 41, table 7); no local norms exist and the foreign norm is fifteen years old, so the score is read as approximate. With a standard error of measurement of 3.4 points, the 95% interval is from about 61 to about 75. The score is therefore compatible with the lower part of the distribution but cannot be reliably separated from the cut-off of 70. No effort test was carried out, and the person was tested earlier by a physician; the earlier score is not available, so a practice effect cannot be excluded. The score describes performance at the time of testing and says nothing about an earlier period. The described difficulty in following complex instructions also appears in the documentation in the file. Whether the person understands the consequences of signing a contract in the legal sense is for the court to decide.

The corrected passage is longer, but every added sentence answers one of the ten questions. That does not make the opinion weaker; it makes it possible for someone else to check it.

For experts: the test section of an opinion

An expert writing a test-based opinion is advised to give the test section the same order every time, which makes it easier for the court to read and harder for the expert to forget an element. The order follows the ten questions.

- **Purpose and choice:** which question in the instruction the test is to clarify and why this instrument was chosen.
- **Instrument:** name, edition, language, source of translation, proof of legitimate use.
- **Evidence of validity and reliability:** sources for this language and population; what is missing.
- **Procedure:** date, duration, conditions, interpreter, deviations.
- **Results:** raw and derived, the norm with its source, the confidence interval.
- **Response checks:** which were carried out, what they showed, what was not checked.
- **Interpretation:** what the score means for the question in the instruction; which other sources were used.
- **Limitations:** the limits of the instrument and of this testing; alternative explanations.
- **Conclusion:** factual language tied to the measurement; legal qualification left to the court.
- **File:** where the raw material is kept and how it can be passed to a qualified person.

An opinion should be as concise and clear as the case permits. That does not conflict with a detailed test section: the measure is that every sentence carries a fact the reader can check. A section that fits on two pages and answers the ten questions is better than one that does not ask them.

Questions a court may put to an expert

The questions are neutral and do not assume that the opinion is wrong. They can supplement the question in the instruction or be used at a hearing.

- Which exact instrument and version was used, and in which language?
- Is there published evidence of its validity for this purpose and this population?
- Against which norm was the score interpreted, and on whom was it collected?
- What is the standard error of measurement, and in what range can the score be expected?
- Where does the cut-off come from, and what does it mean for this person?
- Has the person already been tested with this or a similar test?
- Which checks of response style and effort were carried out, and what did they show?
- Where does the instrument and the key come from, and can the raw material be checked by another qualified expert?
- What does the score not show?
- Which other sources were used, and do they agree with the test?
- Which statements in the opinion concern facts and which concern an assessment of probability?

Five common misunderstandings

"The test is objective, so the score is a fact." Scoring is objective, if done well. The interpretation of a score depends on norm, cut-off, purpose and context, and there is room in it for error and bias.

"A score in the average range proves the person is capable." An average score on one test under controlled conditions does not prove ability in all situations. The same holds in reverse.

"The more precise the scores, the more accurate they are." Decimals do not increase accuracy. Accuracy is limited by the reliability of the test, which is expressed by the standard error of measurement.

"If the test is in a manual, it is valid." A manual describes the evidence for the population on which the test was tried. Whether it applies to this person, in this language and for this purpose has to be shown.

"Failing an effort test means malingering." It means the pattern of performance is not consistent with full effort. The cause may be intent, but also pain, fatigue, misunderstanding or language. A conclusion about intent needs more than one result.

Checklist: the ten questions on one page

Complete the list alongside the opinion. Each "no" or "not stated" is not necessarily an error, but it is a place for a question.

#	Question	Yes	No	Note
1	The instrument, version and language are named, and the source of evidence of validity for this language and purpose is cited			
2	The norm is named (manual, year, sample) and justified for this person; its age is considered			
3	The cut-off has a source and a purpose; how common the property is in the group is considered			

4	Scores are given with the SEM and an interval; classification near a boundary is conditional			
5	Earlier use of the same or a similar test was examined; changes are interpreted with reliability			
6	Response-style and effort checks are stated with their outcome; intent is not inferred from one score			
7	The origin of the instrument and the key is documented; the expert's qualification is clear; access to raw data is possible			
8	There is an express section on limitations; cause, intent and past state are not inferred from the score			
9	Conclusions are factual and tied to a measurement; legal concepts are left to the court			
10	The file allows the scoring to be repeated; the records the rules require were kept			

Glossary

- **Practice effect:** improvement in a score on solving the same or a similar test again, owing to experience with the tasks.
- **Cut-off:** a score that separates two categories; an agreement, not a fact of nature.
- **Confidence interval around a score:** the range in which, with the stated probability, the score can be expected if the same person were tested again.
- **Measurement invariance:** the property of an instrument to measure the same concept in the same way in different groups or languages.
- **Norm:** the distribution of scores of a reference group, against which an individual score is interpreted.
- **Sensitivity and specificity:** the share of people with the property that the cut-off finds, and the share of people without the property that it correctly sets aside.
- **Percentile:** the share of the group whose score is equal or lower.
- **Reliability:** the consistency of scores on repeated measurement.
- **Malingering:** the deliberate presentation of symptoms or weakness for external gain; different from a deviating response pattern.
- **Standard error of measurement (SEM):** the expected deviation of a score owing to chance factors, expressed in the units of the test.
- **Response style:** a tendency towards a particular way of responding (agreeing, exaggerating, socially desirable responding) regardless of the content of the question.
- **Validity:** the degree to which evidence and theory support the interpretation of scores for a particular purpose and population.

About the author and an interest to declare

Dr Milos Kankaras is a psychometrician with a PhD in social sciences (Tilburg University) and more than 25 years of research experience in developing and checking measurement instruments, including large international studies. He does not carry out clinical evaluations of individuals. He develops psychological instruments and provides measurement-quality services, including independent expert opinions on whether assessment instruments are valid, so readers should know of that interest. The guide was prepared as a professional contribution and does not recommend or reject any individual test.

Questions and comments: milos@centerforpsychology.me | <https://miloskankaras.com>

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, 26(4), 543-570. <https://doi.org/10.1080/13854046.2012.680913>
- International Test Commission. (2018). ITC guidelines for translating and adapting tests (second edition). *International Journal of Testing*, 18(2), 101-134. <https://doi.org/10.1080/15305058.2017.1398166>
- Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, 59(1), 12-19. <https://doi.org/10.1037/0022-006X.59.1.12>
- Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society*, 18(4), 625-630. <https://doi.org/10.1017/S1355617712000240>
- Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, 52(3), 194-216. <https://doi.org/10.1037/h0048070>
- Pietschnig, J., & Voracek, M. (2015). One century of global IQ gains: A formal meta-analysis of the Flynn effect (1909-2013). *Perspectives on Psychological Science*, 10(3), 282-306. <https://doi.org/10.1177/1745691615577701>
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71-90. <https://doi.org/10.1016/j.dr.2016.06.004>
- Rogers, R., & Bender, S. D. (Eds.). (2018). *Clinical assessment of malingering and deception* (4th ed.). Guilford Press.