SOUTH AFRICAN CLINICAL NEUROPSYCHOLOGICAL ASSOCIATION

www.sacna.co.za

# UNIFORM LABELLING OF COGNITIVE TEST SCORES: RECOMMENDATIONS FROM SACNA BASED ON GUILMETTE ET AL. (2020)

Urvashi Maganlal, Sharon Truter, Ann B. Shuttleworth-Edwards

## INTRODUCTION

An important theme throughout the article was that interpretations of scores are different from the labelling of scores: scores cannot be impaired, only functions can be impaired. For this reason, test score labels were chosen that do not imply interpretation but only the frequency or commonality of performance, and the word "score" was recommended as part of the test score label in order to distinguish the difference between a particular test result and an ability. To determine if a function is impaired, various factors need to be taken into account besides the test scores. These factors are described in Guilmette et al. (2020) and will not be repeated here.

Guilmette et al. (2020) have provided test score labels for four categories of performance-based tests: tests with normal distributions; tests with non-normal (highly skewed) distribution; tests used to determine the absence or presence of pathognomonic signs; and performance validity tests (PVT's). These four test types are described below, with suggested elaboration by the present authors to facilitate their use.

## 1. Tests with normal distribution

For the tests that have normal distribution (such as IQ tests), the working group decided on a seven-category labelling model, where the average range accounts for 50% of the normal distribution. Using this model with standard scores (where the mean is 100 and standard deviation is 15), the recommended label for a standard score of 90-109 is "average score"; for 110-119 is "high average score", for 120-129 is "above average score" and for 130 and higher is "exceptionally high score". At the other end of the spectrum, the recommended label for a standard score of 80-89 is "low average score", for 70-79 is "below average score" and for lower than 70 is "exceptionally low score". It was suggested that t-scores, z-scores and percentiles follow a similar labelling approach. Accordingly, Dr Sharon Truter has produced a table that can facilitate the use of the standard score and scaled score labelling recommendations of Guilmette et al. (2020) as defined above, for z-scores, t-scores or percentiles (see Table 1, below).

Table 1. Test score labels for tests with *normal distributions* where data are available as standard scores, scaled scores, z-scores, t-scores and percentiles.

| Score Label | Standard Score[1] | Scaled Score | Z-Score | T-Score | Percentile | Percent Included |
|---|---|---|---|---|---|---|
| Exceptionally high score | ≥130 | 16-19 | 2 and above | 70+ | ≥98 | 2.2 |
| Above average score | 120 – 129 | 14,15 | 1.3 to 2 | 63 – 69 | 91-97 | 6.7 |
| High average score | 110 – 119 | 12,13 | 0.6 to 1.3 | 56 – 62 | 75-90 | 16.1 |
| Average score | 90 – 109 | 9-11 | ±0.6 | 44 – 55 | 25-74 | 50.0 |
| Low average score | 80 – 89 | 7,8 | -0.6 to -1.3 | 37 – 43 | 9-24 | 16.1 |
| Below average score | 70 – 79 | 5,6 | -1.3 to -2.0 | 30 – 36 | 2-8 | 6.7 |
| Exceptionally low score | <70 | 1-4 | -2.0 and below | 29 and below | <2 | 2.2 |

*Note.* [1] *"Standard score" here refers to a score where the mean is 100 and standard deviation is 15 as with IQ scores.*
*Source:* Table compiled by Dr Sharon Truter based on the labelling recommendations of Guilmette et al. (2020) in combination with information on the percentage distribution of the population expressed in scaled scores, z-scores, t-scores and percentiles extrapolated from a number of additional sources (Lezak et al., 2012; Mitrushina et al., 2005; Strauss, Sherman & Spreen, 2006).

The z-score is a statistic based on the examinee's raw score in relation to the test's normative mean and standard deviation (SD) (*z-score = examinee's score – normative test mean ÷ normative test SD*). However, some practitioners prefer to use norms reported in means and standard deviations directly, without conducting the extra step of calculating a z-score. In order to do this, the practitioner uses the reported normative test means and standard deviations exactly as they are to establish where an examinee's score falls according to the following divisions: 1 SD above and below the mean; 1 to 2 SDs above and below the mean; and >2 SDs above and below the mean (see Table 2). This is the method employed in the South African norm-based case studies of Shuttleworth-Edwards (2010; 2012; 2016).

If this method is employed, to be broadly compatible with the Guilmette et al. (2020) categories, it is important to note that the 1 SD and 1 to 2 SD categories are wide and require differentiation depending on whether the examinee's score falls within the upper or lower limits of the ranges. Also, as a rule of thumb it is important always to note where a score lies within any category, especially when it closely borders on another category.

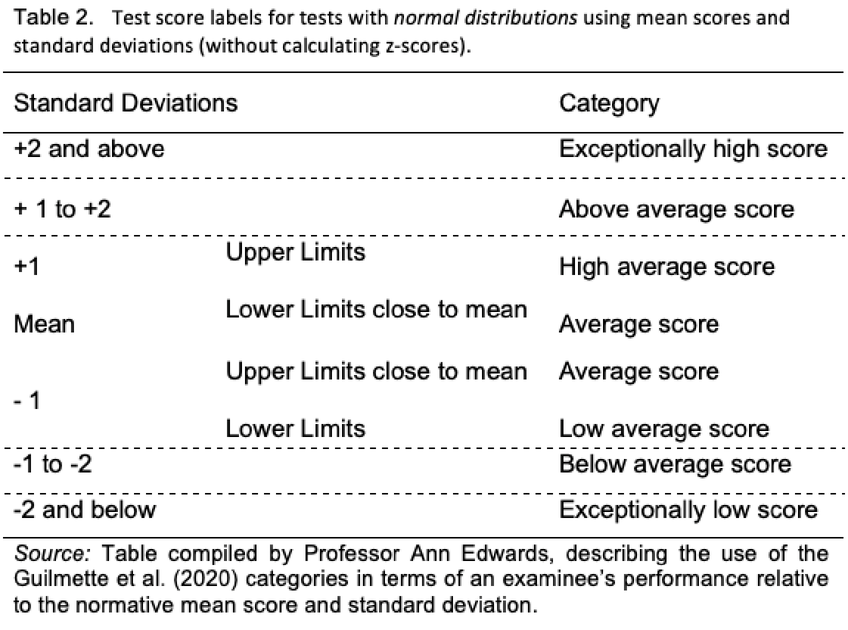

Table 2. Test score labels for tests with *normal distributions* using mean scores and standard deviations (without calculating z-scores).

| Standard Deviations | | Category |
|---|---|---|
| +2 and above | | Exceptionally high score |
| + 1 to +2 | | Above average score |
| +1 | Upper Limits | High average score |
| Mean | Lower Limits close to mean | Average score |
| | Upper Limits close to mean | Average score |
| - 1 | Lower Limits | Low average score |
| -1 to -2 | | Below average score |
| -2 and below | | Exceptionally low score |

*Source:* Table compiled by Professor Ann Edwards, describing the use of the Guilmette et al. (2020) categories in terms of an examinee's performance relative to the normative mean score and standard deviation.

*Finally, be vigilant about the meaning of high versus low scores in terms of ability, when using the mean and SD descriptive statistic directly, or in the form of a z-score. Where high scores reflect good performance (e.g. number of correct responses on a task), this accords with above average categories of performance. On the other hand, where high scores reflect poor performance (e.g. time taken to complete a task, and error scores), then above average scores indicate poor ability. In other words, an exceptionally high score may indicate exceptionally good performance, but in the case of timed scores and error scores an exceptionally high score will indicate exceptionally poor ability.*

## 2. Tests with non-normal (highly skewed) distribution

Guilmette et al. (2020) point out that tests with highly skewed distributions, such as those with ceiling or floor effects, are often designed to identify deficits, not exceptional performance, and therefore labelling higher scores on these tests as “high average score”, “above average score” or “exceptionally high score” may not be meaningful and could be misleading. An example of a test with a skewed distribution is the copy trial of the Rey Complex Figure, where many normal, healthy adults achieve close to perfect scores. Guilmette et al. (2020) recommend that for tests that have highly skewed distribution, test scores that are within or above the average range (above the 24th percentile) should be labelled as “within normal expectations” or “within normal limits”, and that test scores falling below the average range be labelled in accordance with their delineation of test score labels as they relate to percentiles (Guilmette et al. 2020, Table 2), while cautioning that not all normally distributed tests will fit the frame they have provided.

Guilmette et al. (2020) make it clear that standard scores should not be used with tests that are not normally distributed. Rather, percentiles should be used (where they are available) because percentiles are based on actual cumulative counts of individuals who obtained a specific score.  The same argument applies to z-scores, as well as means and standard deviations on which the z-score is based. There is a problem, though, in that the normative data available to practitioners on our commonly employed cognitive tests, are often only provided in the form of means and standard deviations, even though some of these tests may not be normally distributed and have ceiling effects. In such instances, the present authors suggest practitioners label scores that are at the mean and higher as “within normal expectations” or “within normal limits”. With non-normal distributions it is impossible to calculate where scores would fall in relation to percentile ranks if only means and standard deviations are available as normative data. Scores that are well below the mean can broadly be deemed to be poorer than expected, and might be very cautiously further interpreted according to the categories given by Guilmette et al. (2020) in their Table 2.

*Competent neuropsychologists should understand every test they use, know its purpose, and develop an awareness of its score distribution.*

## 3. Tests used to determine the absence or presence of pathognomonic signs

Tests that are used to determine the presence or absence of pathognomonic signs (such as tests of apraxia like the Luria Three-Step Test, also known as Fist/Side/Flat) are generally not affected by demographic variables. For this reason, it was agreed that performance scores for these signs are unnecessary and may even be misleading. Guilmette et al. (2020) recommended that for these tests, the actual pathognomonic sign or specific behaviour observed should be identified and named, and then described using the labels of “intact”, “present” or “absent”.

## 4. Performance validity tests (PVT’s)

For PVTs (such as the Test of Memory Malingering; TOMM) that are used to identify feigned or suspect effort, test engagement and test validity, Guilmette et al. (2020) recommended a three-tier labelling system: “valid range”, “indeterminate range” and “invalid range”.  Other descriptors (such as “pass” or “fail”, “acceptable” or “unacceptable”, or “below chance level of performance”) were rejected based on the lack of specificity or conciseness and the view that these terms were judgemental or confusing. It was highlighted that an invalid score on a PVT does not always indicate the presence of malingering or feigned effort and may or may not invalidate or compromise all test results.

## 5. Concluding Comments

In concluding, Guilmette et al. (2020) assert that the objectives of the position paper were not meant to be instructive or limit interpretations but rather to provide best practice guidelines to clinicians towards a common language. In so doing confusion would be reduced and the reports would be easier to understand by all readers including the referral sources, the legal fraternity, trainees and colleagues. They suggest that for the sake of clarity, the recommended test score labels in their paper be used, irrespective of the score labels given in test manuals. Clinicians are advised by them to look at confidence intervals and error bands for scores on or near cut-off points. They are also advised to include a graph or table that explicitly identifies which standard scores apply to which labels in their reports. To this end, readers are invited to make use of the tables provided in their article.

In the presentation of this summary, SACNA supports the Guilmette et al. position, while including some elaboration on the various test types to facilitate practitioner use.

## REFERENCES

Guilmette, T.J., Sweet, J.J., Hebben, N., Koltai, D., Mahone, E.M., Spiegler, B.J., Stucky. K., Westerveld, M. (2020). Conference Participants American Academy of Clinical Neuropsychology consensus conference statement on uniform labeling of performance test scores. *The Clinical Neuropsychologist. 34*(3),437-453.

Link: https://www.tandfonline.com/doi/pdf/10.1080/13854046.2020.1722244?needAccess=true

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). Neuropsychological Assessment (Vol. V). Oxford: Oxford University Press.

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). Handbook of Normative Data for Neuropsychological Assessment (Second ed.). Oxford: Oxford University Press.

Shuttleworth-Edwards, A.B. (2010). Practitioner guidelines for career counselling in light of cross-cultural influences on WAIS-III IQ test performance. *Journal of Psychology in Africa, 20*(3), 413-419.1

Shuttleworth-Edwards, A.B. (2012). Guidelines for the Use of the WAIS-IV with WAIS-III Cross-cultural Normative Indications**.** *South African Journal of Psychology, Special Issue on Cognitive Science and Neuropsychology in Southern Africa, 42*(3), 2012, 399-410.2

Shuttleworth-Edwards, A.B. (2016). The interpretation of WAIS-IV brain injury test protocols for South African Xhosa speaking individuals using WAIS-III cross-cultural norms. In R. F. Ferraro (Ed.), *Minority and cross-cultural aspects of neuropsychological assessment* (2nd ed., pp. 97-114). New York: Taylor and Francis. 3

Strauss, E. S., Sherman, E. M., & Spreen, O. (2006). A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary. Oxford: Oxford University Press.

**Notes**

This article was first published in Brainwaves, SACNA website, 07 June 2021

Acknowledgements are due to SACNA executive members Dr Frances Hemp and Dr Menachem Mazabow for their editing contributions in the completion of this article.