

American Academy of Clinical Neuropsychology consensus conference statement on uniform labeling of performance test scores

Thomas J. Guilmette, Jerry J. Sweet, Nancy Hebben, Deborah Koltai, E. Mark Mahone, Brenda J. Spiegler, Kirk Stucky, Michael Westerveld & Conference Participants

To cite this article: Thomas J. Guilmette, Jerry J. Sweet, Nancy Hebben, Deborah Koltai, E. Mark Mahone, Brenda J. Spiegler, Kirk Stucky, Michael Westerveld & Conference Participants (2020) American Academy of Clinical Neuropsychology consensus conference statement on uniform labeling of performance test scores, *The Clinical Neuropsychologist*, 34:3, 437-453, DOI: [10.1080/13854046.2020.1722244](https://doi.org/10.1080/13854046.2020.1722244)

To link to this article: <https://doi.org/10.1080/13854046.2020.1722244>



Published online: 10 Feb 2020.



Submit your article to this journal [↗](#)



Article views: 59468



View related articles [↗](#)




View Crossmark data [↗](#)



Citing articles: 130 View citing articles [↗](#)



American Academy of Clinical Neuropsychology consensus conference statement on uniform labeling of performance test scores

Thomas J. Guilmette^a, Jerry J. Sweet^b, Nancy Hebben^{c,d}, Deborah Koltai^e, E. Mark Mahone^f , Brenda J. Spiegler^g, Kirk Stucky^{h,i}, Michael Westerveld^j and Conference Participants^{*}

^aDepartment of Psychology, Providence College; Department of Psychiatry and Human Behavior, Warren Alpert Medical School of Brown University, Providence, RI, USA; ^bDepartment of Psychiatry and Behavioral Sciences, NorthShore University HealthSystem, Evanston, IL, USA; ^cDepartment of Psychiatry, Harvard Medical School, Boston, MA, USA; ^dDepartment of Psychiatry, Cambridge Health Alliance, Cambridge, MA, USA; ^eDepartment of Neurology, Duke University School of Medicine, Durham, NC, USA; ^fDepartment of Neuropsychology, Kennedy Krieger Institute; Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, USA; ^gPrivate Practice, Toronto, ON, Canada; ^hDepartment of Behavioral Health, Hurley Medical Center, Hurley Medical Center, Flint, MI, USA; ⁱDepartment of Medicine, Michigan State University, East Lansing, MI, USA; ^jAdventhealth Neuropsychology-Orlando, Adventhealth Medical Group, Adventhealth Children's Hospital, Orlando, FL, USA

ABSTRACT

Objectives: Descriptive labels of performance test scores are a critical component of communicating outcomes of neuropsychological and psychological evaluations. Yet, no universally accepted system exists for assigning qualitative descriptors to scores in specific ranges. In addition, the definition and use of the term “impairment” lacks specificity and consensus. Consequently, test score labels and the denotation of impairment are inconsistently applied by clinicians, creating confusion among consumers of neuropsychological services, including referral sources, trainees, colleagues, and the judicial system. To reduce this confusion, experts in clinical and forensic neuropsychological and psychological assessment convened in a consensus conference at the 2018 Annual Meeting of the American Academy of Clinical Neuropsychology (AACN). The goals of the consensus conference were to recommend (1) a system of qualitative labels to describe results from performance-based tests with normal and non-normal distributions and (2) a definition of impairment and its application in individual case determinations.

Results: The goals of the consensus conference were met resulting in specific recommendations for the application of uniform labels for performance tests and for the definition of impairment, which are described in this paper. In addition, included in this

ARTICLE HISTORY

Received 2 January 2020
Accepted 10 January 2020
Published online 10 February 2020

KEYWORDS

Impairment; performance tests; qualitative descriptors; test score labels

CONTACT Thomas J. Guilmette  tguilmet@providence.edu  Providence College, Providence, RI 02918, USA

***Conference Participants (by working group):** Normal Distribution Group: Nancy Hebben & E. Mark Mahone (Co-chairs), Corwin Boake, Desiree Byrd, Jennifer Haut, Jennifer Koop, David J. Schretlen. Non-Normal Distribution Group: Kirk Stucky & Brenda J. Spiegler (Co-chairs), Gordon Chelune, Veronica Bordes Edgar, Daryl Fujii, Joseph Snow, Jerry J. Sweet. Impairment Definition and Application Group: Deborah Koltai & Michael Westerveld (Co-chairs), Laurence Binder, Leigh D. Hagan, Laura Kenealy, Bernice Marcopulos, June Yu Paltzer.

consensus statement is a description of the conference process and the rationales for these recommendations.

Conclusions/Importance: This consensus conference is the first formal attempt by the professional neuropsychological community to make recommendations for uniform performance test score labels and to advance a consistent definition of impairment. Using uniform descriptors and terms will reduce confusion and enhance report comprehensibility by the consumers of our reports as well as our trainees and colleagues.

Statement of the problem

The most common method of describing test score performance in both clinical and forensic neuropsychological reports is by the use of qualitative descriptors (Guilmette, Hagan, & Giuliano, 2008), such as average, above average, superior, and impaired, which are generally regarded as more meaningful and comprehensible than raw scores, standard scores, or percentile ranks in communicating results. Whereas multiple labeling systems have been proposed (e.g. Groth-Marnat, 2009; Heaton, Grant, & Matthews, 1991; Schoenberg & Rum, 2017; Schretlen, Testa, & Pearlson, 2010; Wechsler, 2009, 2014), no consensus or universally accepted system exists for assigning qualitative descriptors or labels for performance-based tests.

Twenty-five years ago, in his presidential address to Division 22 (Rehabilitation Psychology) of the American Psychological Association, Bruce Caplan identified the problem by stating “Terms such as ‘moderately impaired’ and ‘within normal limits’ frequently lack quantitative referents and thus are subject to differing interpretations across individuals and contexts” (1995, p. 236). Caplan’s concern grew out of a study he conducted in which he asked participants of a “major neuropsychological organization” to assign a descriptive label from a list of 22 potential ratings to four hypothetical patients. He found a remarkable degree of inter-rater inconsistency. Caplan further opined that “In order to increase our credibility, especially in forensic contexts where different clinicians may provide disparate interpretations of identical test scores, we need research and discussion toward some consensus on what various descriptive labels imply” (p. 239). Similarly, Hebben and Milberg (2002) in their review of neuropsychological test interpretation also concluded, “Labels such as ‘average’ or ‘below average’ are not precise and may refer to different score ranges depending on the individual clinician.”

In a survey of board-certified neuropsychologists, Guilmette et al. (2008) asked respondents to assign a descriptive label to 12 different standard scores from 50 to 130 derived from a memory test based on a brief case scenario. The mean number of *different* descriptive labels assigned by the 110 survey participants to each of the 12 standard scores was 14 with a range of 9 to 23. This result again provided empirical evidence of the significant variability and lack of uniformity in assigning qualitative descriptors to specific standard scores. Importantly, Guilmette and colleagues also found label assignment variance related to different interpretive methods or standards. Some respondents relied on normative or inter-individual comparative descriptors (e.g. superior, average, below/above average, low), on impairment or intra-individual

comparisons that describe normality or abnormality based on some estimated premorbid baseline (e.g. normal, intact, impaired), or on intra-individual comparisons regarding the expectation of a score based on other factors such as IQ that were included in the case scenario (e.g. below/above expectations). Given the lack of reliability in applying performance test descriptors, Guilmette et al. suggested that “professional neuropsychological and clinical psychology endeavor to articulate specific recommendations or adopt explicit standards that establish well-defined terminology and more consistent assignment of qualitative descriptions for test score ranges” (2008, p. 136).

Adding to the confusion for clinicians in deciding which test score labels to apply, test publishers are inconsistent in their recommendations for descriptors of the scores for their tests. As trainees and practicing clinicians are well aware, different test manuals may recommend different qualitative labels for the same standard scores. For example, the Wechsler intelligence and memory test manuals list qualitative descriptors for their standard scores and most clinicians apply those descriptors when describing performance on those scales. However, the qualitative descriptor in the Wechsler system for a specific standardized score is not always the same descriptor that is recommended for the same score on a different test. Consider a standard score of 75, which would be labeled “borderline” by the adult Wechsler tests, “very low” by the Wide Range Achievement Test-5 (Wilkinson & Robertson, 2007) and the Wechsler Intelligence Scale for Children-5 (Wechsler, 2014), “well below average” by the Kaufman Short Neuropsychological Assessment Procedure (Kaufman & Kaufman, 1994), “low” by the Woodcock-Johnson IV (Schrank, McGrew, & Mather, 2014), “poor” by the Test of Nonverbal Intelligence-3 (Brown, Sherbenou, & Johnsen, 1997), and “below average to mildly impaired” by the Boston Qualitative Scoring System for the Rey-Osterrieth Complex Figure (Stern et al., 1999). Consequently, the clinician is likely inclined to assign different labels to the same standard score from different tests to adhere to the test publisher recommendations. While deviations from such varied test manual recommendations may be quite justified, such deviations may render clinical interpretations vulnerable to attack in litigious contexts. These inconsistencies in a report can be confusing to both patients and referral sources, as well as for trainees trying to understand the complexities of test score interpretation.

Similarly, identification of an “impaired” test score range has been treated inconsistently by researchers and clinicians, with cutoffs variably applied to scores beginning at one standard deviation, 1.5 standard deviations, or two standard deviations below normative expectations (Beauchamp et al., 2015; Heaton et al., 1991; Ingraham & Aiken, 1996; Meyer, Boscardin, Kwasa, & Price, 2013; Schoenberg et al., 2018). In the Guilmette et al. (2008) survey noted previously, “impaired” was applied by some respondents as a descriptive label beginning with a standard score as high as 95. Not surprisingly, the term impairment (along with other terms such as defective, abnormal, and deficient) was applied with increasing frequency as standard scores declined.

When classifying scores as impaired, disagreement has also occurred regarding the labels that identify degree of abnormal performance, sometimes beginning with a term conveying uncertainty (e.g. borderline) before proceeding to using labels that convey apparent greater certainty (e.g. mild, moderate, and severe) for scores that presumably deviate to a greater extent from normative expectations. However, these

modifiers also lack any consensus definition. In each of these instances, the use of terms meant to convey abnormality or “impairment” has typically been based purely on a test score deviating from normative expectations. This practice habit can be thought of as ‘test bound,’ an inappropriate concrete approach that considers each specific test score as having inherent clinical meaning, without considering the overall test result profile and the particular examinee’s life context. Such an approach is not viewed as an acceptable method of arriving at clinical conclusions.

The lack of consistency in applying test score labels and of a definition of the term impairment undercuts the professional practice of clinical neuropsychology. Despite being a decades-old issue, it has not been adequately addressed by our professional organizations. To this end, the American Academy of Clinical Neuropsychology (AACN) established a consensus conference to propose a consistent set of qualitative descriptors and to define impairment with the goal of providing guidance to clinicians and facilitating increased consistency in the application of these terms.

Development of the consensus conference

The genesis of this consensus conference began in February 2014 when the AACN Board of Directors (BOD) approved a proposal initiated by Manfred Greiffenstein, with assistance from Thomas Guilmette, to form a workgroup to create a position paper for guidance on the uniform labeling of test scores. The Board entrusted Greiffenstein and Guilmette to select workgroup members, develop key survey questions, organize and analyze the survey data collection among the members, and write a position paper. The workgroup co-chairs asked 13 expert neuropsychologists/psychologists who represented researchers, clinicians and opinion leaders to participate in this endeavor with the provision that each member could agree that (a) labeling guidelines were necessary, (b) level of performance and score interpretation needed to be distinguished, and ultimately that (c) score interpretation should always be contextualized. The members of this original workgroup included: Corwin Boake, Bruce Caplan, Robert Denney, Jacobus Donders, Anthony Giuliano, Leigh Hagan, Bernice Marcopulos, Ann Marcotte, Scott Millis, Neil Pliskin, Kirk Stucky, Joseph Snow, and Keith Owen Yeates.

A series of online surveys and suggestions were distributed among the work group members with subsequent communication and discussion occurring via email. The goal was not consensus, but rather, to find a majority opinion regarding descriptors for test scores with normal distributions and non-normal distributions as well as a definition of impairment. Progress was slow and incomplete but a preliminary draft of the position paper was written by Greiffenstein and Guilmette and forwarded to the AACN Publications Committee for review in June 2016. However, work on the project ceased following the death of Manfred Greiffenstein in August 2016.

Guilmette then consulted with the chair of the Publications Committee of AACN, Jerry Sweet, who suggested that the contents of the position paper draft likely would not be instructive to clinicians, and particularly given the importance of this issue, a consensus conference would be the most appropriate means of advancing the use of uniform test labels of performance tests. Guilmette and Sweet approached the AACN Board of Directors with a proposal for a consensus conference modeled after the

AACN consensus conference on the neuropsychological assessment of effort, response bias, and malingering (Heilbronner et al., 2009). The AACN BOD approved the proposal at its annual meeting in June 2017. The all-day consensus conference was planned to occur during the 2018 annual meeting to be held in San Diego. The AACN BOD also approved that approximately 25 conference attendees, with relevant expertise and experience, would be invited by the co-chairs, Guilmette and Sweet, to propose labels for normally distributed scores, non-normally distributed scores, and to define impairment and its use and application. All potential conference members would also need to agree on the three stipulations set forth by Greiffenstein and Guilmette for their position paper proposal noted above. Following the consensus conference, a 90-minute presentation or “public forum” would be held and open to all AACN conference attendees. The public forum would include presentation of the consensus conference recommendations and elicit critical feedback from attendees. A consensus conference statement, summarizing the results and recommendations, would then be submitted to AACN’s official journal, *The Clinical Neuropsychologist*, for publication.

Conference organizers identified 28 experts from Canada and the United States who represented diversity across gender, adult/pediatric focus, culture, and work setting to participate in the conference. Five experts interested in participating in this process were not available to attend the conference, and as an alternative agreed to provide reviews of the consensus statement prior to publication although only two eventually were able to review the manuscript. The remaining conference members, based in part on preference and the diversity considerations, were divided into three work groups that would seek consensus on test score labels for normally distributed tests, non-normally distributed tests, and impairment definition and application. Work group co-chairs were identified, again reflecting a balance of gender and adult versus pediatric foci. Pre-conference readings and references, which were identified by the organizers and from suggestions made by the conference members themselves, were distributed to all the participants. Resource materials included scientific or scholarly articles and book chapters (listed in the [appendix](#)) as well as other informal test score labeling proposals, including those from the initial Greiffenstein and Guilmette labeling workgroup, and an abbreviated version of the International Classification of Functioning from the World Health Organization.

Process of creating consensus and writing the consensus statement

Conference participants met on June 20, 2018, the day before the annual meeting of the American Academy of Clinical Neuropsychology in San Diego. Of the 25 members who were scheduled to meet, three participants were unable to attend but agreed to review the consensus statement prior to publication, resulting in a total of 22 attendees from 17 states in the US, the District of Columbia, and Canada. Following a brief overview of the history and goals of the conference, participants assembled in their work groups and began discussion. The remainder of the day alternated between work group breakouts and overall group discussion in attempts to reach consensus in the three domains. The entire group reached consensus on a number of specific points in all three work group areas. Subgroup co-chairs then summarized their findings and

recommendations in a presentation to AACN conference attendees during a 90-minute public forum on June 22, 2018 during which feedback and criticism were solicited.

To encourage transparency and openness to the views of the AACN membership at large, the PowerPoint presentation from the June 22, 2018 meeting was distributed via the AACN listserv, prompting comments, recommendations, critiques, and questions from July 11 to July 28, 2018. The comments and feedback received from the listserv resulted in additional e-mail discussion and consideration among the original 22 consensus conference participants. Essentially, all initial consensus recommendations that came out of the June 20, 2018 conference were reconsidered.

The significant online discussion and consideration of multiple alternatives led to the recognition that the initial points of consensus attained during the June meeting had evolved, leading to a need to re-establish consensus. Work group co-chairs and conference co-chairs worked with each group, with revised points of consensus brought to the larger group for final consideration. Eventually, consensus was again reached for all three subgroup topics: test score labels for tests with non-normal distributions on December 13, 2018, for impairment definition and application on February 1, 2019, and for test score labels for tests with normal distributions on May 2, 2019.

A summary of the consensus process and its recommendations was written and reviewed by all 22 conference participants and was also reviewed by outside experts, resulting in this consensus conference statement. The consensus statement reflects the combined expertise of 27 neuropsychologists/psychologists, who considered scientific literature, historical perspectives, and clinical factors, as well as extensive input from many AACN members. This input and review also included the AACN Publication Committee and ultimate approval by the AACN Board of Directors.

The recommendations contained in this statement should not be considered mandatory practice standards. Rather, they reflect consensual expert guidance or “best practices” that clinicians can consider incorporating into their work to achieve greater uniformity and consistency in the application of test score labels or descriptors and the use of the term impairment. To be absolutely clear, this statement is *not* intended to instruct or limit clinicians in their interpretation of neuropsychological test data. The integrative analysis of a neuropsychological test profile rests solely with the judgment of individual clinicians and their appreciation for and expertise in synthesizing information from multiple medical, historical, cultural, behavioral, and other sources to arrive at clinical formulations, impressions, and diagnoses.

Consensus recommendations for test labels for tests with normal distributions

The normal distribution work group initially relied on the following concepts to guide its deliberations:

- Interpretation of scores is different from labeling of scores.
- Scores cannot be “impaired;” only a function can be impaired.
- Simplicity of descriptors can improve communication.
- Descriptors should be based on the frequency or commonality of performance, not on pathology.

In considering the best way to anchor labels to test scores, 5-category and 7-category models were reviewed for their relative merits. The 5-category model would have labels assigned to scores at each standard deviation. For example, ± 1 standard deviation would be encompassed under one label, average, with other labels assigned for each additional standard deviation above and below the mean up to three standard deviations, yielding five test score labels overall. The primary advantages of this approach would be that labels easily map onto standard deviations and there is simplicity in having relatively few labels to consider. However, this model diverges meaningfully from common clinical practices, making adoption by practitioners less likely. In addition, the average range would extend across two standard deviations encompassing about 68% of the distribution. In contrast, when standard scores between 90 and 109 are labeled as average, about 50% of the distribution falls within this range, which is consistent with most descriptive systems (Groth-Marnat, 2009; Schoenberg & Rum, 2017; Schretlen et al., 2010; Wechsler, 2009). Consequently, the 7-category model derived from the Wechsler system was adopted. This model has more clinical relevance, with finer gradations that are not linked to integer standard deviation units. In addition, this model was considered closer to current clinical practices, and thus likely easier for clinicians to incorporate into their practices.

As was true within the consensus group, the specific labels assigned to various score ranges were a matter of detailed discussion during the feedback session with the AACN conference attendees and, subsequently, among the AACN listserv members who offered comments and suggestions. There was a strong belief that test score labels should be free of terms that appear judgmental, biased, or would be viewed as representing a clinical conclusion, and, instead, should reflect only a score position within the normal distribution. Specifically, the intent was that score labels not appear to convey the separate process of clinician interpretation, which is the necessary step in determination of impairment or deficit.

The initial test score labels recommended by the consensus conference were as follows: extremely high score (≥ 130); high score (120–129); above average score (110–119); average score (90–109); below average score (80–89); low score (70–79); and extremely low score (< 70). There were initial concerns raised at the open AACN meeting that the term “extremely” did not adequately reflect the uncommon frequency of test scores at the very ends of the distribution. Following discussion among the consensus conference members, it was agreed to change the term to “exceptionally.” Notably, the most prolonged and detailed debate and consideration involved the labels “low average” or “below average” in the 80–89 standard score range due to the potential ambiguity of whether “low average” is still considered “average” and whether a standard score below 90 should be considered “below average.” One consideration was that up to 24% of the population would be considered “below average” if the cutoff for “average” was all standard scores below 90. This concern appeared to be particularly salient among some of the pediatric neuropsychologists. Discussion was held regarding the use of an additional modifier, such as “slightly” below or “mildly” below average, but these terms were rejected due to their ambiguity and lack of standard meaning. In trying to find a resolution between these considerations and given the acceptance of the “low average” label

for standard scores 80–89 among many clinicians and extant qualitative descriptor systems (Groth-Marnat, 2009; Schretlen et al., 2010; Wechsler, 2009), the consensus panel recommended the “low average” descriptor for standard scores between 80 and 89.

Given that standard scores between 80 and 89 are labeled “low average,” then scores falling below that level, in the 70–79 standard score range, are considered “below average.” The adult Wechsler classification system refers to this range as “borderline,” but that term was considered too ambiguous and prone to imply an interpretive conclusion. As noted previously, other test publishers have described scores in this range as “low,” “very low,” “well below average,” “poor,” and “below average to *mildly impaired*.” These terms were also rejected on grounds of appearing to be judgmental, biased, too open to interpretation, or conflate a test score with an impairment label. Although the original consensus conference recommended “low score” as a label in the 70–79 standard score range as opposed to the current “below average,” this was ultimately rejected because this range reflects scores that are unambiguously below average and, as indicated above, fall below scores that are “low average.” Also, the description of a score being simply “low” or “high” (in the 120–129 range) was believed to be too ambiguous and open to interpretation. Last, suggestions that the modifier “well” be added to the labels “below average” and “above average” were also regarded by a majority of the group co-chairs as adding little, if any clarifying value.

The final consensus recommendations for descriptive labels for normally distributed test scores is listed below in the context of general standard scores commonly used in intelligence tests. Transforming other types of scores, such as T-scores, z-scores, or percentiles, into qualitative descriptors would follow the same labeling approach. Whereas, with most performance-based tests, lower standard scores indicate worse performance, in select instances, higher standard scores can indicate worse performance. In these instances, clinicians choose labels that reflect this distinction (Table 1).

The consensus group also recommends that clinicians specify the normative group and any demographic adjustments used for standard score determination (e.g. if scores are adjusted for sex, age, education, etc.). Clinicians should also recognize that nomenclature is based on specific derived scores, which themselves are psychometric estimates bounded by confidence intervals. Thus, clinicians should give careful consideration to labeling scores near cut-points, including consideration of the error band. In addition, the group consensus is that this system be used instead of those provided within specific test manuals, as this will promote discipline-wide uniformity and facilitate consistent and effective communication with stakeholders. Finally, to clarify further the assignment of labels and descriptors to test scores, a consensus recommendation is that clinicians

Table 1. Recommended test score labels based on standard scores and percentiles for tests with normal distributions.

Standard Score	Percentile	Score Label
>130	>98	Exceptionally high score
120–129	91–97	Above average score
110–119	75–90	High average score
90–109	25–74	Average score
80–89	9–24	Low average score
70–79	2–8	Below average score
<70	<2	Exceptionally low score

include a table or graph within reports that explicitly identifies which standard scores coincide with which labels. This is especially important as we recognize that, despite our efforts, the lay public and other consumers may have difficulty appreciating and understanding the distinctions among our recommended qualitative test score labels.

As noted previously, these test score labels are intended solely to be descriptive, identifying positions of scores relative to a normal curve distribution. As such, the labels do not convey impairment or other evaluative judgments; scores in isolation cannot be impaired or deficient. Acknowledging that risk of a score representing an impaired function increases with statistical deviation from normative expectations, nevertheless, there is consensus that the determination of deficits or impairments is the responsibility of the clinician, who arrives at such a determination using a broad range of information specific to the individual patient. Consistent with this intent, in describing test scores, the consensus recommendation was to place the word “score” following the descriptor, in order to emphasize the difference between a specific test result and an ability. As a practical matter, clinicians may find it cumbersome to always place the word “score” following the descriptor (e.g. low average score), and thus may decide to drop the word to decrease redundancy and enhance conciseness.

Consensus recommendations for test labels for tests with non-normal distributions

In clinical neuropsychology four types of tests are frequently administered that have non-normal distributions.

- a. Tests intended to assess specific cognitive domains but with highly skewed distributions in the normal population (e.g. Boston Naming Test, Judgment of Line Orientation (JLO), clock drawing, figure copy, etc.).
- b. Tests used to determine the presence or absence of pathognomonic signs or specific conditions (e.g. tests for apraxia, manual motor sequencing, sensory-perceptual exam, etc.).
- c. Performance validity tests (PVTs) and measures primarily used to identify concerns regarding test engagement, symptom magnification, effort, and test validity (e.g. Test of Memory Malingering, Word Memory Test, Advanced Clinical Solutions Word Choice, etc.).
- d. Questionnaires and rating scales regarding cognitive abilities and/or behavioral conditions or symptoms frequently assessed by neuropsychologists (e.g. Behavioral Rating Inventory of Executive Function, Behavioral Assessment System for Children, Child Behavior Checklist, etc.).

The consensus conference participants did not address test score labeling for questionnaires and rating scales in Group D because these are not performance-based tests; providing recommendations on score labeling for these types of instruments was beyond the scope of our AACN BOD mandate. For tests in Groups A, B, and C, the purpose of the test administration and the type of information the test provides are fundamentally different from one another, as well as from tests that have normal or

near-normal score distributions. Consequently, we addressed tests in each category separately and have provided a summary discussion with recommendations in the relevant sections below.

A. Tests with highly skewed distributions

Tests in this category are fundamental components of a thorough neuropsychological assessment, as many are designed to assess a specific cognitive ability or domain (e.g. Judgment of Line Orientation, Neuropsychological Assessment Battery Naming Test, categories completed on the Wisconsin Card Sorting Test [WCST], recognition testing within the Brief Visuospatial Memory Test – Revised [BVM-T-R], etc.). The nature of such tests is more comparable to criterion-based measures that assess a specific ability for which there is little variability among individuals considered “normal” or healthy. In general, the purpose of these tests is to identify specific areas of impairment or deficit in examinees, unlike normative measures that show high variability with “normal” or healthy individuals and situate results within the normal population distribution. Two organizing questions related to these measures were contemplated by group members in depth: Is it appropriate to use standard scores for tests with highly restricted ranges? and Should scores on these tests be labeled differently than tests with normal distributions?

With regard to tests with highly restricted score ranges, the consensus was that percentiles should be used instead of standard scores. The rationale for this recommendation is based on the fact that percentile ranks are more comparable and meaningful than other transformed scores when the distribution is highly skewed. Importantly, the percentiles for non-normally distributed tests are based on actual cumulative counts of individuals who obtained a specific score and thus are not statistical estimations based on standard deviation units around the mean of the reference group. Thus, we recommend avoiding the use of standard scores for these test results.

For some tests with skewed distributions, normality can be approximated via various “smoothing” procedures. The use of standard scores in these situations may be justifiable, but the clinician should carefully weigh the risks and benefits of standard score transformation and give additional consideration to how those scores should be labeled.

Regarding the question of whether scores on these tests should be labeled differently than tests with normal distributions, the group consensus was that the labels should be the same between the two types of tests, for the following reasons:

- Using a common language and simplified system for descriptive labels across the two types of tests is much less confusing for clinicians and consumers.
- The use of a separate labeling system for tests in this category would create an unnecessarily complex system that could be difficult to employ in some clinical settings.
- At times, the neuropsychologist may not know whether the underlying distribution for a specific test is normal or non-normal. Additionally, the test’s underlying distribution might be subject to change depending on specific demographic variables (e.g. sex, age, education, and multicultural considerations).
- Competent neuropsychologists should understand the test, its purpose, and its score distribution in the normal population.

This recommendation to apply comparable descriptive labels to tests with normal or skewed distributions is made with four important exceptions:

- a. The same labels used with normally distributed tests are recommended, with the qualification that percentile rank should be used to determine the label, not a standard score. This is straightforward when applied to labeling scores at the lower end of the distribution, but not the upper end of the distribution (see b. below).
- b. On highly skewed tests, it is sometimes statistically impossible to attain a percentile score in the higher ranges. On many such tests a perfect or near perfect raw score is typically described as being at or above the 16th percentile. For example, a perfect score of 6 categories correct on the WCST is noted as simply above the 16th percentile. This is also true for a perfect score on BVMT-R recognition or Rey Complex Figure Copy. Considering this measurement and statistical reality, describing such scores as anything but being *within normal expectations* or *within normal limits* would be inappropriate.
- c. Given that skewed tests have significant ceiling or floor effects and are often designed to identify deficits, not exceptional performance, labeling higher scores on these tests as above average or exceptionally high (even when the percentile range is high) may not be meaningful and could be misleading. For example, the JLO, a 30-item test, has a low ceiling in that 28% of the normative sample earned corrected scores of 29–30 and scores above 21 were earned by 93% of the sample. Although scores of 29 or 30 fall at the 86th percentile, labeling these scores as superior, as classified in the manual (p. 59), is not as meaningful as simply indicating that the score was within normal limits or within normal expectations (Benton, Sivan, Hamsher, Varney, & Spreen, 1994). To elaborate this point, a score at the 86th percentile on the JLO does not have the same clinical meaning as a test with normally distributed scores, such as Block Design, when scoring at the 86th percentile. For the latter, high percentile ranks always indicate that a small percentage of the normative sample earned a score in that upper range. As this example illustrates, this is not true for tests with highly skewed distributions. Thus, we recommend that practitioners refrain from using the descriptors high average, above average, or exceptionally high when labeling scores at the upper end of a highly skewed distribution. Rather, using a descriptive label conveying the general meaning of a test score, such as performance was *within normal expectations* or *within normal limits*, would be more appropriate, including test scores that fall within the average range, or above the 24th percentile. The table below elucidates non-normal distribution test score recommendations based on percentiles. We caution, however, that not all non-normally distributed tests will fit the example we have provided. Importantly, these labels should not to be applied to PVTs (see letter C below).
- d. Finally, for tests in which smoothing procedures have been employed in norms development, the use of the “exceptionally high” score label is strongly discouraged because this label is descriptively reserved for tests with genuine normal or near-normal distributions, namely when scores in the exceptionally high category represent performances at or above the 98th percentile (Table 2).

Table 2. Recommended test score labels based on percentiles for tests with non-normal distributions.

Percentile	Score Label
>24	Within Normal Expectations Score or Within Normal Limits Score
9–24	Low Average Score
2–8	Below Average Score
<2	Exceptionally Low Score

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

Tests to determine the presence or absence of pathognomonic signs or specific conditions are not typically impacted by various demographic variables. For example, healthy adults would be expected to perform with few to no errors or irregularities on tests of praxis, motor sequencing, and line bisection. Of course, exceptions arise in children because of developmental considerations. For example, certain language-based errors exhibited by a 4-year-old are normal, but if observed in a 17-year-old such errors would be considered pathognomonic (e.g. letter reversals, frequent paraphasic errors, etc.). The primary questions we considered were, “When specific disorders, syndromes, or pathognomonic signs are apparent during testing is reporting a score necessary, useful, or accurate? Might doing so be misleading?”

After careful consideration, the consensus recommendation is that when an examinee exhibits a specific pathognomonic sign or neurobehavioral condition, it should be named and/or described in specific behavioral terms. For example, “On various language-based tasks, speech was non-fluent with multiple paraphasic errors. He was unable to read or write. However, he could repeat words and short sentences. These findings are consistent with transcortical motor aphasia.” Also, when referring to negative findings or the absence of pathology or abnormal performance on these types of tests the use of terms *intact*, *present*, or *absent* is suggested, as appropriate to the type of sign. Our position is that describing or naming a pathognomonic sign or condition is much more informative and accurate than assigning a score even if score ranges are available. Use of labels for test scores in this category of tests is not as meaningful or informative as specific and precise descriptions of the performance or identification of the specific condition/syndrome. A competent neuropsychologist has a sophisticated understanding of brain-behavior relationships and will be skilled at identifying classic neurobehavioral presentations without need for test scores (e.g. aphasia, apraxia, hemispatial inattention, agnosia, etc.). This is particularly apparent when assessment procedures are used to reveal or investigate the presence or absence of pathognomonic signs or specific neurobehavioral conditions in category B.

C. Performance validity tests

Of all the test types considered by the non-normal distribution work group, this one received the most attention from the neuropsychological community. The reasons for this are likely multifactorial, but certainly linked to the implications of labeling scores in a specific fashion, especially in forensic contexts.

After considering a number of suggestions provided by AACN clinicians interested in this particular category of tests, the consensus was that the following three-tiered system for labeling scores should be used – *valid range*, *indeterminate range*, *invalid range*.

Although a variety of existing systems were considered, many were rejected because they contained or implied an interpretive position (e.g. pass vs. fail), were potentially judgmental (acceptable vs. unacceptable), lacked specificity or conciseness, or did not adequately capture the range of reasons for low performance (i.e. an individual can obtain low scores on PVTs for various reasons, intentional withholding of effort being one).

The possible inclusion of a fourth category label (i.e. *below chance level performance*), was discussed, but was rejected for the following reasons: (1) moving beyond the label of invalid range score into a subset of scores below chance level may appear to move beyond description to interpretation within the overall invalid range; (2) adding a subset score range within the range already labeled invalid would potentially be more confusing and more difficult to apply consistently across practitioners; and (3) a competent neuropsychologist is expected to comment on significantly below chance level performances and implications when integrating all pertinent information in their interpretive summary and case formulation.

A critical point is that attaining an invalid range score on a PVT does not always or automatically indicate the presence of malingering or “compromised effort” and may or may not invalidate all testing results. With regard to such issues, the current consensus conference participants had no areas of disagreement with the practice recommendations outlined in the 2009 AACN consensus statement on response validity and malingering (Heilbronner et al. 2009). In situations in which an examinee produces one or more invalid range or indeterminate range score(s), it is ultimately the clinician who is responsible for judging, based on the totality of information available, what those scores mean and how they should be interpreted.

Finally, examples of how these labels might be used in a report are provided in three separate examples below. These examples might be included in a report section that describes individual test results. In each example there is clear reference to a score, rather than a specific interpretive statement.

- The examinee’s score on a stand-alone PVT was within the valid range.
- On an embedded PVT, the examinee attained an indeterminate range score.
- The forced-choice memory PVT score was within the invalid range.

Consensus recommendations for impairment definition and application

Of the three areas considered by the consensus conference, the impairment definition evoked the fewest number of responses and suggestions from the greater AACN neuropsychological community. However, following the consensus conference itself and the posting of our recommendations on the AACN listserv, the consensus conference participants revisited the initial definition of impairment. Ongoing dialogue and discussion resulted in the following consensus recommendation for the term impairment:

Neuropsychological impairment is abnormal neurocognitive or neurobehavioral capacity. Impairment may result from loss of previously acquired skill or result from atypical

development, may be transient or fixed across time, and can have variable impact on functional capacity and disability. Test scores, per se, do not define impairment. A combination of factors, including test scores that deviate from expectations, and other findings related to functional capacity, identify neuropsychological impairment.

In applying the impairment definition in individual cases, the following factors, among others, should be considered.

- Normal intra-individual variability and the frequency of low scores in normal populations (Binder, Iverson, & Brooks, 2009; Donnell, Belanger, & Vanderploeg, 2011; Heyanka, Holster, & Golden, 2013; Palmer, Boone, Lesser, & Wohl, 1998; Schretlen, Munro, Anthony, & Pearlson, 2003). The latter is related, among other factors, to the number of tests administered and the cut-point used to define abnormality.
- The convergence of shared versus unique variance among tests.
- The characteristics of the normative/comparison standard (e.g. demographically stratified versus general population norms).
- Performance validity.
- Test engagement.
- Cultural factors associated with different diversities (e.g. language, literacy, level and quality of education, familiarity and comfort with the testing situation, testing biases, opportunities for learning, conception of intelligent behavior, and communication style).
- Emotional and medical conditions, medications, physical (non-illness) and cognitive factors.
- High scores, or the lack of low scores, do not preclude the determination of functional limitations or “impairment.” Conversely, low scores do not necessarily indicate functional impairment; consideration of context is required to make such determinations.
- The functional relevance of the finding in the context of the referral.
- Environmental and task demands as well as supports that ameliorate or mitigate the neurocognitive or neurobehavioral capacity.

In reporting results to referral sources, information should be clear and specific, and convey meaningful interpretive conclusions, such as indicating the presence or absence of impairment, or that findings are equivocal. This reporting can be made for individual domains or overall functioning.

Summary

The lack of uniformity in the application of performance test score labels has been a longstanding problem in clinical neuropsychology. This consensus conference is the first formal attempt by the professional neuropsychological community to make recommendations for uniform performance test score labels and to advance a consistent definition of impairment. Our recommendations are not mandates or standards, but rather, represent expert consensus opinion on these important issues. We hope that clinicians will incorporate our recommendations into their clinical practices to increase the uniformity of test score descriptors, the most frequent way in which test performance is communicated in clinical and forensic reports. Using uniform descriptors and

terms will reduce confusion and enhance report comprehensibility by the consumers of our reports as well as our trainees and colleagues.

Our recommendations are in no way meant to interfere with or restrict the *interpretation* of test scores, which continues to rest solely on the clinical judgment of the professional. We recognize and accept that not all neuropsychologists will find our recommendations appropriate for adoption in their clinical practices or will agree with our consensus recommendations. Nonetheless, our consensus recommendations are the first organized attempt by our specialty to attain test descriptor uniformity and, as such, may initiate an ongoing specialty-wide dialogue about this critical issue. We also recognize that our recommendations are not fixed in stone and that the introduction of new assessment methods and technologies may require future modifications to accommodate those innovations. Consequently, the consensus conference participants respectfully welcome continued dialogue to further develop and refine our nomenclature. We also wish to acknowledge the time and effort of all the members of the AACN neuropsychological community who contributed useful feedback and suggestions to assist us in this worthwhile endeavor.

Acknowledgements

The authors wish to express their gratitude to the following external reviewers: Robert L. Denney, Jacobus Donders, Anthony J. Giuliano, Mike R. Schoenberg, and Keith Owen Yeates, as well as input and approval from the AACN Publication Committee and AACN Board of Directors.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

E. Mark Mahone  <http://orcid.org/0000-0002-5022-1499>

References

- Beauchamp, M. H., Brooks, B. L., Barrowman, N., Aglipay, M., Keightley, M., Anderson, P., ... Zemek, R. (2015). Empirical derivation and validation of a clinical case definition for neuropsychological impairment in children and adolescents. *Journal of the International Neuropsychological Society*, 21(8), 596–609. doi:10.1017/S1355617715000636
- Benton, A. L., Sivan, A. B., Hamsher, K., Varney, R. R., & Spreen, O. (1994). *Contributions to neuropsychological assessment: A clinical manual* (2nd ed.). New York, NY: Oxford University Press.
- Binder, L. M., Iverson, G. L., & Brooks, B. L. (2009). To err is human: "Abnormal" neuropsychological scores and variability are common in healthy adults. *Archives of Clinical Neuropsychology*, 24(1), 31–46. doi:10.1093/arclin/acn001
- Brown, L., Sherbenou, R.J., & Johnsen, S.K. (1997). *Examiner's manual test of nonverbal intelligence* (3rd ed.). Austin, TX: Pro-Ed.
- Caplan, B. (1995). Choose your words. *Rehabilitation Psychology*, 40(3), 233–240. doi:10.1037/h0092829
- Donnell, A. J., Belanger, H. G., & Vanderploeg, R. S. (2011). Implications of psychometric measurement for neuropsychological interpretation. *The Clinical Neuropsychologist*, 25(7), 1097–1118. doi:10.1080/13854046.2011.599819

- Groth-Marnat, G. (2009). *Handbook of psychological assessment* (5th ed.). Hoboken, NJ: John Wiley and Sons, Inc.
- Guilmette, T. J., Hagan, L., & Giuliano, A. J. (2008). Assigning qualitative descriptors to test scores in neuropsychology: Forensic implications. *The Clinical Neuropsychologist*, 22(1), 122–139. doi: [10.1080/13854040601064559](https://doi.org/10.1080/13854040601064559)
- Heaton, R. K., Grant, I., & Matthews, C. G. (1991). *Comprehensive norms for an expanded Halstead-Reitan Battery: Demographic corrections, research findings, and clinical applications*. Odessa, FL: Psychological Assessment Resources.
- Hebben, N., & Milberg, W. (2002). *Essentials of neuropsychological assessment*. New York, NY: John Wiley and Sons.
- Heilbronner, R. L., Sweet, J. J., Morgan, J. E., Larrabee, G. J., Millis, S. R., & Conference Participants. (2009). American Academy of Clinical Neuropsychology consensus conference statement on the neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist*, 23, 1093–1129. doi: [10.1080/13854040903155063](https://doi.org/10.1080/13854040903155063)
- Heyanka, D. J., Holster, J. L., & Golden, C. J. (2013). Intraindividual neuropsychological test variability in healthy individuals with high average intelligence and educational attainment. *International Journal of Neuroscience*, 123(8), 526–531. doi: [10.3109/00207454.2013.771261](https://doi.org/10.3109/00207454.2013.771261)
- Ingraham, L. J., & Aiken, C. B. (1996). An empirical approach to determining criteria for abnormality in test batteries with multiple measures. *Neuropsychology*, 10(1), 120–124. doi: [10.1037/0894-4105.10.1.120](https://doi.org/10.1037/0894-4105.10.1.120)
- Kaufman, A. S., & Kaufman, N. L. (1994). *Kaufman Short Neuropsychological Assessment Procedure manual*. Bloomington, MN: Pearson.
- Meyer, A.-C. L., Boscardin, W. J., Kwasa, J. K., & Price, R. W. (2013). Is it time to rethink how neuropsychological tests are used to diagnose mild forms of HIV-associated neurocognitive disorders? Impact of false-positive rates on prevalence and power. *Neuroepidemiology*, 41(3-4), 208–216. doi: [10.1159/000354629](https://doi.org/10.1159/000354629)
- Palmer, B. W., Boone, K. B., Lesser, I. M., & Wohl, M. A. (1998). Base rates of “impaired” neuropsychological performance among healthy older adults. *Archives of Clinical Neuropsychology*, 13, 503–511. doi: [10.1016/S0887-6177\(97\)00037-1](https://doi.org/10.1016/S0887-6177(97)00037-1)
- Schoenberg, M. R., Osborn, K. E., Mahone, E. M., Feigon, M., Roth, R. M., & Pliskin, N. H. (2018). Physician preferences to communicate neuropsychological results: Comparison of qualitative descriptors and a proposal to reduce communication errors. *Archives of Clinical Neuropsychology*, 31, 631–643. doi: [10.1093/arclin/acx106](https://doi.org/10.1093/arclin/acx106)
- Schoenberg, M. R., & Rum, R. S. (2017). Towards reporting standards for neuropsychological study results: A proposal to minimize communication errors with standardized qualitative descriptors for normalized test scores. *Clinical Neurology and Neurosurgery*, 162, 72–79. doi: [10.1016/j.clineuro.2017.07.010](https://doi.org/10.1016/j.clineuro.2017.07.010)
- Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock-Johnson IV tests of cognitive abilities*. Rolling Meadows, IL: Riverside.
- Schretlen, D. J., Munro, C. A., Anthony, J. C., & Pearlson, G. D. (2003). Examining the range of normal intraindividual variability in neuropsychological test performance. *Journal of the International Neuropsychological Society*, 9(6), 864–870. doi: [10.1017/S1355617703960061](https://doi.org/10.1017/S1355617703960061)
- Schretlen, D. J., Testa, S. M., & Pearlson, G. D. (2010). *Calibrated Neuropsychological Normative System professional manual*. Lutz, FL: Psychological Assessment Resources.
- Stern, R. A., Javorsky, D. J., Singer, E. A., Singer Harris, N. G., Somerville, J. A., Duke, L. M., ... Kaplan, E. (1999). *The Boston qualitative scoring system for the Rey-Osterrieth figure*. Lutz, FL: Psychological Assessment Resources.
- Wechsler, D. (2009). *WMS-IV technical and interpretive manual*. San Antonio, TX: Pearson.
- Wechsler, D. (2014). *Wechsler intelligence scale for children* (5th ed.). San Antonio, TX: Pearson.
- Wilkinson, G. S., & Robertson, G. J. (2007). *WRAT-5 Wide Range Achievement Test 5th Edition professional manual*. Bloomington, MN: Pearson.

Appendix

- Beauchamp, M. H., Brooks, B. L., Barrowman, N., Aglipay, M., Keightley, M., Anderson, P., ... Zemek, R. (2015). Empirical derivation and validation of a clinical case definition for neuropsychological impairment in children and adolescents. *Journal of the International Neuropsychological Society*, 21(8), 596–609. doi:[10.1017/S1355617715000636](https://doi.org/10.1017/S1355617715000636)
- Brooks, B. L., & Iverson, G. L. (2012). Improving accuracy when identifying cognitive impairment in pediatric neuropsychological assessments. In E. Sherman & B. Brooks (Eds.), *Pediatric forensic neuropsychology* (pp. 66–88). New York, NY: Oxford University Press.
- Busch, R. M., Chelune, G. J., & Suchy, Y. (2006). Using norms in neuropsychological assessment of the elderly. In D. Attix & K. Welsh-Bohmer (Eds.), *Geriatric neuropsychology assessment and intervention* (pp. 133–157). New York, NY: The Guilford Press.
- Chelune, G. J., & Duff, K. (2013). The assessment of change: Serial assessments in dementia evaluations. In L. D. Ravdin & H. L. Katzen (Eds.), *Handbook on the neuropsychology of aging and dementia, clinical handbooks in neuropsychology* (pp. 43–57). New York, NY: Springer Science + Business Media, LLC.
- Erodi, L. A., & Lichtenstein, J. D. (2017). Invalid before impaired: An emerging paradox of embedded validity indicators. *The Clinical Neuropsychologist*, 31, 1029–1046. doi:[10.1080/13854046.2017.1323119](https://doi.org/10.1080/13854046.2017.1323119)
- Guilmette, T. J., Hagan, L., & Giuliano, A. J. (2008). Assigning qualitative descriptors to test scores in neuropsychology: Forensic implications. *The Clinical Neuropsychologist*, 22(1), 122–139. doi:[10.1080/13854040601064559](https://doi.org/10.1080/13854040601064559)
- Ingraham, L. J., & Aiken, C. B. (1996). An empirical approach to determining criteria for abnormality in test batteries with multiple measures. *Neuropsychology*, 10(1), 120–124. doi:[10.1037/0894-4105.10.1.120](https://doi.org/10.1037/0894-4105.10.1.120)
- Meyer, A.-C. L., Boscardin, W. J., Kwasa, J. K., & Price, R. W. (2013). Is it time to rethink how neuropsychological tests are used to diagnose mild forms of HIV-associated neurocognitive disorders? Impact of false-positive rates on prevalence and power. *Neuroepidemiology*, 41(3–4), 208–216. doi:[10.1159/000354629](https://doi.org/10.1159/000354629)
- Schoenberg, M. R., Osborn, K. E., Mahone, E. M., Feigon, M., Roth, R. M., & Pliskin, N. H. (2018). Physician preferences to communicate neuropsychological results: Comparison of qualitative descriptors and a proposal to reduce communication errors. *Archives of Clinical Neuropsychology*, 31, 631–643. doi:[10.1093/arclin/acx106](https://doi.org/10.1093/arclin/acx106)
- Schoenberg, M. R., & Rum, R. S. (2017). Towards reporting standards for neuropsychological study results: A proposal to minimize communication errors with standardized qualitative descriptors for normalized test scores. *Clinical Neurology and Neuropsychology*, 162, 72–79. doi:[10.1016/j.clineuro.2017.07.010](https://doi.org/10.1016/j.clineuro.2017.07.010)