Testing the limits

Cautions and concerns regarding the new Wechsler IQ and Memory scales

David W. Loring, PhD
Russell M. Bauer, PhD

Address correspondence and reprint requests to Dr. David W. Loring, Department of Neurology, Emory University, 101 Woodruff Circle, Suite 6000, Atlanta, GA 30322
dloring@emory.edu

ABSTRACT

The Wechsler Adult Intelligence Scale (WAIS) and the Wechsler Memory Scale (WMS) are 2 of the most common psychological tests used in clinical care and research in neurology. Newly revised versions of both instruments (WAIS-IV and WMS-IV) have recently been published and are increasingly being adopted by the neuropsychology community. There have been significant changes in the structure and content of both scales, leading to the potential for inaccurate patient classification if algorithms developed using their predecessors are employed. There are presently insufficient clinical data in neurologic populations to insure their appropriate application to neuropsychological evaluations. We provide a perspective on these important new neuropsychological instruments, comment on the pressures to adopt these tests in the absence of an appropriate evidence base supporting their incremental validity, and describe the potential negative impact on both patient care and continuing research applications. *Neurology*® 2010;74:685–690

GLOSSARY

FSIQ = full-scale IQ; **GAI** = General Ability Index; **PIQ** = performance IQ; **PRI** = Perceptual Reasoning Index; **RCI** = reliable change index; **VCI** = Verbal Comprehension Index; **VIQ** = verbal IQ; **WAIS** = Wechsler Adult Intelligence Scale; **WMS** = Wechsler Memory Scale.

Characterizing cognitive abilities is an important part of the comprehensive neurologic workup in many patient populations (e.g., dementia, traumatic brain injury, movement disorders, epilepsy, multiple sclerosis). In these settings, neuropsychological performance is used to identify procedure-related risk factors, cognitive disease effects, or to measure the benefits or adverse events associated with various therapies. Neuropsychologists rely heavily on standardized measures of IQ and memory when making their diagnostic formulations.

Like new software releases, psychological test revisions purportedly offer important improvements over previous versions. Test revisions permit modification of test material to allow for content updating to reflect current models of cognitive function, to improve psychometric test properties and operating characteristics, or simply to make test administration and scoring easier. Test revision also insures that there has been no upward drift in test scores over time, and that "average" test performance across the population for tests such as the Wechsler scales remains at 100.[1]

In North America, the most common IQ and memory tests are the Wechsler Adult Intelligence Scale (WAIS) and Wechsler Memory Scale (WMS).[2] Recent revisions of these popular instruments (i.e., WAIS-IV/WMS-IV) were published in 2008/2009, and are increasingly being adopted. As with previous WAIS/WMS revisions, there have been substantial changes including scale indices and subtest content and administration.

Revisions of these scales have important effects on test usage and applicability in clinical and research settings. We describe how the recently revised WAIS and WMS may impact users and consumers of these instruments.

WECHSLER ADULT INTELLIGENCE SCALE–FOURTH EDITION Decades of research have demonstrated that the verbal IQ (VIQ) and performance IQ (PIQ) scores derived from the Wechsler IQ scales are

From the Department of Neurology (D.W.L.), Emory University, Atlanta, GA; and Department of Clinical and Health Psychology (R.M.B.), University of Florida, Gainesville.

Disclosure: Author disclosures are provided at the end of the article.

not homogeneous measures of verbal and nonverbal "intelligence" as originally conceptualized, but instead are comprised of subtests clustering into 4 distinct cognitive domains (verbal comprehension, perceptual reasoning, working memory, processing speed). VIQ not only contains measures assessing verbal abstraction (e.g., Similarities) and knowledge (e.g., Information), but also subtests of attention and working memory (e.g., Digit Span). PIQ reflects visual spatial problem solving ability (e.g., Block Design), but also assesses possessing speed (e.g., Digit Symbol). Because of the consistency of the 4-factor structure identified across multiple studies, formal factor-based composite scores were introduced in the WAIS-III, although traditional VIQ and PIQ scores could still be calculated.

With the release of the WAIS-IV, however, VIQ and PIQ scores have been eliminated entirely. Thus, short of reporting the full-scale IQ (FSIQ), the only way to present summary scores across multiple subtests is to use composite scales. To facilitate interpretation of the revised scale, psychologists are advised that "*the terms (Verbal Comprehension Index-VCI) and (Perceptual Reasoning Index-PRI) should be substituted for the terms VIQ and PIQ in clinical decision-making and other situations where VIQ and PIQ were previously used*"[3] (p. 9, italics in original).

This recommendation, however, is premature. WAIS-IV composite scores reflect more narrow and, in the case of the Verbal Comprehension Index (VCI) and Perceptual Reasoning Index (PRI), less neuropsychologically sensitive measures than VIQ or PIQ to brain impairment, at least in the context of nonfocal brain disease. The WAIS-IV also introduces a new composite score, the General Ability Index (GAI), which like VCI and PRI excludes the working memory and processing speed index in its calculation, both of which together contribute 40% of the variance in FSIQ (p. 170).[4] Thus, the GAI and FSIQ differ from each other in the same way that the VCI and PRI differ from VIQ and PIQ, yet the test publisher states explicitly "the GAI does not replace the FSIQ." Therefore, the VCI and PRI should not be treated as synonymous with VIQ and PIQ in clinical decision-making, and it is clear that using VCI–PRI discrepancy scores for neuropsychological inference will lead to different conclusions than those based upon VIQ vs PIQ discrepancies.

The composition of WAIS-IV subtests has also changed, with some subtests eliminated, others modified, and several completely new subtests introduced (table). The emphasis on rapid solution (speeded performance) has been decreased, with the number of time bonus points on several subtests reduced or eliminated. Since psychomotor slowing is a core feature of many forms of brain injury, the WAIS-IV should be expected to yield fewer FSIQ scores of 70 or below in neurologic populations compared to its predecessor, and will decrease the number of individuals qualifying for special education or disability services using FSIQ cutoff criteria. This also creates a mismatch when historical groups tested on previous Wechsler versions are compared, and will lead to differences in epidemiologic estimates describing prevalence of individuals with FSIQs in the impaired range.

The new WAIS-IV FSIQ has the same overall connotation as the FSIQ of earlier versions, but nevertheless reflects a different composition of cognitive abilities (and thus means something different) than its predecessor. Consequently, the WAIS-IV FSIQ potentially will have altered sensitivity to neuropsychological impairment compared to the WAIS-III FSIQ. Presently, there are insufficient data to establish how new WAIS-IV subtests are affected by various neurologic lesions, and how reconfigured old subtests differ from their predecessors with respect to sensitivity and specificity to neurologic disease.

One of the strongest motivations cited for test revision is to adjust the population average score back to 100, which is necessary since for reasons not fully understood, the average IQ score tends to increase over time (the so-called "Flynn effect"). However, a comparison between WAIS-III and WAIS-IV results suggests only minimal recalibration. The "average" WAIS-III FSIQ compared to the "average" WAIS-IV FSIQ is only 2.9 points higher (table 5.5 in WAIS-IV manual),[3] which most clinicians would

Table WAIS-IV modifications (from table 2.4 in the WAIS-IV manual)

	New test	Administration	Scoring	New items
Block Design		X	X	X
Similarities			X	X
Digit Span		X	X	X
Matrix Reasoning		X	X	X
Vocabulary			X	X
Arithmetic		X	X	X
Symbol Search		X	X	X
Visual Puzzles	X			
Information			X	X
Coding			X	X
Letter-Number Sequencing		X	X	X
Figure Weights	X			
Comprehension			X	X
Cancellation	X			
Picture Completion			X	X

Abbreviation: WAIS = Wechsler Adult Intelligence Scale.

not consider to be clinically significant and which only slightly exceeds the standard error of measurement of 2.3 points for the WAIS-III.[5] On the individual subtest level, scaled score equivalents changed at least 1 point on 9 of 11 subtests, and 4 changed 1.8 points comparing the WAIS to the WAIS-R.[6] However, comparing the WAIS-III to the WAIS-IV, the greatest individual subtest change on the measures common to both versions was 1.0 point and was observed for only a single subtest (Vocabulary) (p. 75).[3] Unlike the WAIS-III, the WAIS-IV standardization sample carefully excluded older normative participants to insure that cases with mild dementia were not accidently included. The WAIS-IV also explicitly excluded subjects from the normative data who demonstrated poor effort or inadequate task engagement during standardization. Rather than reflecting a gradual improvement in test performance over time, this small increase in FSIQ between the WAIS-III and WAIS-IV may simply reflect a normative sample with a greater proportion of healthy individuals and who expended adequate effort during the standardization process.

WECHSLER MEMORY SCALE–FOURTH EDITION Unlike the WAIS-IV, ample research evidence identified important limitations of the WMS-III that could be addressed during the test revision. Each WMS edition has introduced new subtests and new subtest combinations. Unfortunately, with each subsequent release, many of the previously "new" subtests have been discarded due to their failure to adequately assess their intended memory constructs, and replaced with another set of "improved" memory measures. Thus, each subsequent WMS revision fails to provide an accumulated corpus of subtests, and instead represents frequent midstream changes in test development.

The most significant WMS improvement, which was introduced informally to the original 1945 scale, was the inclusion of a 30-minute delayed recall component to the Logical Memory and Visual Reproduction subtests,[7] which in the original version consisted of immediate recall only, although major limitations remained.[8,9] When the scale was subsequently formally revised, significant problems were quickly identified with the WMS-R,[10] although the increased normative information was a significant improvement. The WMS-III addressed many WMS-R concerns, but clinical experience after its release failed to replicate the factor structure reported at publication.[11] In addition, replacing one of the subtests first introduced with the WMS-III (Faces) with one that previously had "optional status" (Visual Reproduction) resulted in a decline in the magnitude of discrepancy scores required to infer probable impairment.[12]

There is insufficient justification for some of the subtest changes incorporated in the WMS-IV. For example, Logical Memory is a measure of prose passage recall using 2 different short story paragraphs, and has been included in various forms with all WMS editions. In the WMS-III, there are 2 learning trials for 1 of the paragraphs, although each story is presented only once in the WMS-IV. No justification or rationale for this change is provided, and this modification in administration may reduce its sensitivity, since story repetition affects its attentional loading. As observed by Jones-Gotman et al.,[13] "part of the reason why some authors have not observed material-specific lateralization effects in clinical studies has been their use of single-trial memory tasks as opposed to tasks emphasizing learning over trials." There is sufficient experience with WMS-III Logical Memory in many clinical samples over the decade since its publication so that the decision to retain or discard multiple Logical Memory trials should have been made empirically.

Some WMS-IV changes may ultimately prove to be of clinical benefit. A new visual memory test (Designs) assesses memory for visual images within a grid, requiring the examinee to recall both visual and spatial information. Whether this subtest becomes a clinically usefully visual memory measure or simply becomes just one more visual memory test in a long line of measures with insufficient sensitivity and specificity for routine use[14] remains to be determined. The WMS-IV also now includes a modified and shortened set of tests for use with older patients (65 and older) to decrease the assessment burden in this population.

The WMS-IV has relaxed the scoring criteria for Visual Reproduction, deemphasizing drawing accuracy and placing greater emphasis on memory, an approach that has been applied to other visual memory tasks such as the Complex Figure.[15] Historically, however, other "improvements" associated with WMS revision, such as nonverbal paired associates for the WMS-R, have been dropped from subsequent revisions since research after test release failed to support their clinical utility.[4]

DISCUSSION Although any new drug or device cannot be approved without appropriate research prior to its availability to clinicians, no such expectation exists for psychological tests. Standards exist for test construction and internal psychometric characteristics, but not for establishing the test's utility in differential diagnosis or its performance in evaluating the target populations with which it will be used.

These tests often form the basis for important clinical decisions such as establishing neurosurgical candidacy (e.g., epilepsy surgery, DBS), predicting risk of postoperative cognitive decline, or appropriateness of educational or institutional disability accommodations (e.g., whether additional time should be allowed for standardized testing for students with processing speed deficits), yet there are no available data on how the new scales will perform in these contexts.

The importance of test validity cannot be overstated, and both WAIS-IV and WMS-IV manuals contain multiple correlation tables establishing their psychometric validity based upon factor-analytic studies and relationships to other cognitive/neuropsychological measures. However, critical information on criterion validity referenced to clinical benchmarks, which reflects how well test results predict a certain diagnosis or functional outcome and which is of central importance to neuropsychological applications of the Wechsler scales, is absent. Thus, important neuropsychological test characteristics addressing criterion validity such as sensitivity, specificity, receiver operating characteristics, or multiple level likelihood ratios cannot be calculated.[16-19]

Prior to market introduction, drug or device manufacturers must rigorously demonstrate the efficacy and safety of their products for their intended clinical application. Psychological test publishers are not required to adhere to a similar principle. The absence of clinical data to guide interpretation is readily acknowledged in the manual: "the data from these samples are presented as examples and are not intended to be fully representative of these diagnostic groups"[4] (p. 105). Indeed, the clinical samples are surprisingly small. For example, only 8 patients who had undergone left anterior temporal lobectomy were included in the clinical samples, and no preoperative epilepsy surgery candidates were present, precisely the group in whom diagnostic sensitivity and specificity would be most helpful.

Changing the structure of neuropsychological tests with each revision impedes the accumulated development of a clinical knowledge base and adversely affects long-term research/database implementation, and there are multiple examples of resistance to such changes in interdisciplinary research settings. In order to maintain fidelity with longitudinal studies, the Uniform Data Set for AD Centers includes components of the WMS-R and WAIS-R, although this limits generalizability to contemporary neuropsychological assessment. The MATRICS Consensus Cognitive Battery,[20] designed for clinical trials in schizophrenia, includes the Hopkins Verbal Learning Test–Revised) which, unlike the WMS, is unlikely to change in structure or administration. In making recommendations for neuropsychological test usage in NIH epilepsy trials, the neuropsychology common data elements subgroup selected the Rey Auditory Verbal Learning Test as their recommended verbal memory test because of concern about changes in stimuli and administration associated with WMS revisions.

Test revisions also take one of the most useful metrics for evaluating neuropsychological change over time out of the hands of the neuropsychologists. Neuropsychology has increasingly relied on reliable change indices (RCIs) in both clinical practice and research to characterize significant cognitive change.[21,22] RCIs reflect whether performance change upon retesting exceeds what can statistically be attributed to test-retest variability, standard error of the test, and practice effects, thereby permitting the determination of significant change at the individual patient level.[23,24] RCI tables exist for many neuropsychological tests including the WAIS-III and WMS-III,[25-27] although it will likely require many years before similar RCI tables can be generated for the WAIS-IV and WMS-IV. This problem would not exist if the appropriate test-retest research had been conducted prior to test publication.

One might argue that if the revised Wechsler scales have such drawbacks, clinicians could simply choose not to use them until an adequate database on criterion validity and clinical utility accumulates. Prevailing ethical standards (the 2002 American Psychological Association Ethical Guidelines and Code of Conduct), however, explicitly state that psychologists should not use "outdated" assessment instruments, and to do so is an enforceable violation that has been exercised by some state Psychology Boards. Several states and agencies have explicitly adopted the new tests as the version required for establishing disability and applying for Americans with Disabilities Act–related accommodations.

What makes a test "outdated?" In our view, the answer should lie in empirical evidence that there is a better, more valid, or more reliable method. In current practice, however, a test becomes outdated when a test publisher releases a new test version based on internal corporate decisions, even when little or no information on test efficacy, clinical utility, or criterion validity accompanies the new release.

When a new neuropsychological test is introduced, those adopting it face significant financial, conceptual, ethical, and clinical challenges. Neuropsychologists who have been trained to base clinical decisions on previous versions of the tests are faced with demands to use a less proven and potentially less effective measure, and with the task of recalibrating

and redeveloping reasonable clinical decision-making algorithms. Given the potential adverse effects on patient care from inaccurate recommendations based upon incompletely validated tests, it is time that more rigorous standards be required to guide the development, validation, and eventual clinical implementation of new psychological and neuropsychological tests. Such standards will insure that test publishers demonstrate incremental validity of their new products prior to marketing and distribution. This is not to say that the new Wechsler scales cannot or will not eventually meet this criterion. However, until they do, there is no reason to disavow the legitimacy of previous test versions. Like Windows computer users who prefer to use the Windows XP operating system over the newer versions (e.g., Vista), we believe that it is appropriate for clinicians and researchers to choose older tests with established clinical utility until there is ample clinical evidence that the newer measures improve diagnostic accuracy and clinical decision-making with appropriate patient populations.

ACKNOWLEDGMENT

The authors thank Pearson/PsychCorp for providing copies of the WAIS-IV and WMS-IV for review.

DISCLOSURE

Dr. Loring serves on scientific advisory boards for the Epilepsy Foundation and Sanofi-Aventis; serves as a Consulting Editor for the *Journal of Clinical and Experimental Neuropsychology, Epilepsy & Behavior,* and *Epilepsy Research*, as a contributing editor for *Epilepsy Currents*, and on the editorial board of *Neuropsychology Review*; has received honoraria for non-industry-sponsored lectures; serves as a consultant for NeuroPace, Inc. and UCB; receives royalties from the publication of *Neuropsychological Assessment, 4th ed.* (Oxford University Press, 2004) and *INS Dictionary of Neuropsychology* (Oxford University Press, 1999); estimates that 50% of his clinical effort involves neuropsychological testing; and receives research support from NeuroPace, Inc., SAM Technology Inc., Myriad Pharmaceuticals, Inc., Novartis, the NIH (NINDS R01038455 [Co-I] and NINDS R01NS031966 [Consultant]), and from the Epilepsy Foundation. Dr. Bauer has received travel expenses and/or honoraria for lectures or educational activities not funded by industry; serves as Co-Editor of *The Clinical Neuropsychologist* and on the editorial board of *Neuropsychology*; receives research support from the NIH (R21 MH64161 [Co-I]); and serves as Co-Director of the Tracking and Evaluation Program, UF Clinical Translational Research Institute (NIH-CTSA).

Received September 25, 2009. Accepted in final form December 3, 2009.

REFERENCES

1. Flynn JR. Massive IQ gains in 14 nations: what IQ tests really measure. Psych Bull 1987;101:171–191.
2. Rabin LA, Barr WB, Burton LA. Assessment practices of clinical neuropsychologists in the United States and Canada: a survey of INS, NAN, and APA Division 40 members. Arch Clin Neuropsychol 2005;20:33–65.
3. Wechsler D, Coalson DL, Raiford SE. Wechsler Adult Intelligence Test: Fourth Edition Technical and Interpretive Manual. San Antonio: Pearson; 2008.
4. Wechsler D, Holdnack JA, Drozdick LW. Wechsler Memory Scale: Fourth Edition Technical and Interpretive Manual. San Antonio: Pearson; 2009.
5. Tulsky DS, Zhu J, Ledbetter MF. WAIS-III WMS-III Technical Manual. San Antonio: The Psychological Corporation; 1997.
6. Lezak MD, Howieson DB, Loring DW. Neuropsychological Assessment, 4th ed. New York: Oxford University Press; 2004.
7. Russell EW. A multiple scoring method for the assessment of complex memory functions. J Con Clin Psychol 1975; 43:800–809.
8. Prigatano GP. Wechsler Memory Scale is a poor screening test for brain dysfunction. J Clin Psychol 1977;33:772–777.
9. Loring DW, Papanicolaou AC. Memory assessment in neuropsychology: theoretical considerations and practical utility. J Clin Exp Neuropsychol 1987;9:340–358.
10. Loring DW. The Wechsler Memory Scale–Revised, or the Wechsler Memory Scale–Revisited? Clin Neuropsychol 1989;3:59–69.
11. Millis SR, Malina AC, Bowers DA, Ricker JH. Confirmatory factor analysis of the Wechsler Memory Scale–III. J Clin Exp Neuropsychol 1999;21:87–93.
12. Hawkins KA, Tulsky DS. Replacement of the Faces subtest by Visual Reproductions within Wechsler Memory Scale–Third Edition (WMS-III) visual memory indexes: implications for discrepancy analysis. J Clin Exp Neuropsychol 2004;26:498–510.
13. Jones-Gotman M, Zatorre RJ, Olivier A, et al. Learning and retention of words and designs following excision from medial or lateral temporal-lobe structures. Neuropsychologia 1997;35:963–973.
14. Barr WB, Chelune GJ, Hermann BP, et al. The use of figural reproduction tests as measures of nonverbal memory in epilepsy surgery candidates. J Int Neuropsychol Soc 1997;3:435–443.
15. Loring DW, Martin RC, Meador KJ, Lee GP. Psychometric construction of the Rey-Osterrieth complex figure: methodological considerations and interrater reliability. Arch Clin Neuropsychol 1990;5:1–14.
16. Ivnik RJ, Smith GE, Cerhan JH, et al. Understanding the diagnostic capabilities of cognitive tests. Clin Neuropsychol 2001;15:114–124.
17. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. Lancet 2005;365:1500–1505.
18. Bowden SC, Loring DW. The diagnostic utility of multiple-level likelihood ratios. J Int Neuropsychol Soc 2009;15:769–776.
19. Barr WB. Receiver operating characteristic curve analysis of Wechsler Memory Scale–Revised scores in epilepsy surgery candidates. Psych Assess 1997;9:171–176.
20. Nuechterlein KH, Green MF, Kern RS, et al. The MATRICS Consensus Cognitive Battery, part 1: test selection, reliability, and validity. Am J Psychiatry 2008;165:203–213.
21. Tröster AI, Woods SP, Morgan EE. Assessing cognitive change in Parkinson's disease: development of practice effect-corrected reliable change indices. Arch Clin Neuropsychol 2007;22:711–718.
22. Hermann BP, Seidenberg M, Schoenfeld J, et al. Empirical techniques for determining the reliability, magnitude, and pattern of neuropsychological change after epilepsy surgery. Epilepsia 1996;37:942–950.
23. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol 1991;59:12–19.

24. Chelune GJ, Naugle RI, Lüders H, et al. Individual change after epilepsy surgery: practice effects and base-rate information. Neuropsychology 1993;7:41–52.
25. Martin R, Sawrie S, Gilliam F, et al. Determining reliable cognitive change after epilepsy surgery: development of reliable change indices and standardized regression-based change norms for the WMS-III and WAIS-III. Epilepsia 2002;43:1551–1558.
26. Vearncombe KJ, Rolfe M, Wright M, et al. Predictors of cognitive decline after chemotherapy in breast cancer patients. J Int Neuropsychol Soc 2009;15:951–962.
27. Heaton RK, Temkin N, Dikmen S, et al. Detecting change: a comparison of three neuropsychological methods, using normal and clinical samples. Arch Clin Neuropsychol 2001;16:75–91.