Contemp School Psychol (2017) 21:255–266
DOI 10.1007/s40688-016-0114-x

SYSTEMATIC REVIEW

Nonverbal and Language-Reduced Measures of Cognitive Ability: a Review and Evaluation

Daniel D. Drevon[1] · Rachel M. Knight[2] · Sharon Bradley-Johnson[1]

Published online: 8 December 2016
© California Association of School Psychologists 2016

Abstract With the number of new and revised nonverbal and language-reduced tests of cognitive ability, selection and interpretation of appropriate measures can be complicated. Seven nonverbal or language-reduced tests with normative data collected within the last 15 years were evaluated. Besides evaluating technical adequacy, other variables affecting test selection and interpretation including adequacy of floors and item gradients, provision of data for subgroups of students, percent of timed items, and response mode are described. Eight additional tests with language-reduced components also were reviewed. Implications for practice and research are presented.

Keywords Assessment · Intelligence · Nonverbal measures · Language-reduced measures · Cognitive assessment

Despite the changing landscape of the field of school psychology, tests of cognitive ability continue to play a major role in psychoeducational assessment (Flanagan et al. 2008). Results from these tests inform high-stakes decisions such as eligibility for special education services and types of services available to students with disabilities. As noted by Reynolds et al. (2006), these tests provide objective evaluations of students' ability which are preferable to subjective opinions of others whose perspectives might be influenced by irrelevant factors.

Though tests of cognitive ability ordinarily rely on verbal interactions between examiner and student to assess ability, these tests are unsuitable for those who have difficulty communicating in or do not communicate in Standard English. These students include those with speech and language disorders, hearing impairments, traumatic brain injury, autism spectrum disorder (ASD), and those who are English language learners (ELLs). For these students, many tests of cognitive ability would yield underestimates of ability because of reliance on verbal interactions. Consequently, a nonverbal test of cognitive ability test may yield fairer and more valid results (McCallum 2003).

Nonverbal tests of cognitive ability measure general ability and are characterized by administration procedures and content that eliminate or reduce the receptive and expressive language required of the student (McCallum 2003; Naglieri and Otero 2012). McCallum (2003) noted that current use of the term *nonverbal* is confusing when applied to tests of cognitive ability because some tests described as *nonverbal* have verbal directions. Bracken and McCallum (1998) suggested that a nonverbal assessment is a process in which neither receptive nor expressive language requirements are placed on the examinee or the examiner. The majority of tests of cognitive ability said to be nonverbal do not meet these criteria. These latter tests, McCallum (2003) suggested, are more appropriately termed *language-reduced* measures. These tests reduce expressive language requirements because students can respond to items by pointing, manipulating objects or materials, or both, eliminating the need for speech. However, test directions are given orally by the examiner which requires understanding of oral language. Some students have difficulty expressing themselves verbally for various reasons (e.g., articulation problems, shy toddlers or preschoolers, students with hearing

✉ Daniel D. Drevon
drevo1dd@cmich.edu

[1] Department of Psychology, Central Michigan University, Mt. Pleasant, MI 48859, USA

[2] Department of Pediatrics, University of Michigan Health Systems, 1011 Cornwell Pl, Ann Arbor, MI 48104, USA

impairments) but understand English better than they are able to orally or manually communicate. For these students, language-reduced measures can be an appropriate option.

However, if students have verbal expression problems, some also may have language comprehension problems which could confound results from language-reduced tests. Consequently, when a language-reduced test is selected, it is incumbent on examiners to ensure their examinees understand the oral directions or directions given via sign language. If it is unclear whether a student understands the directions after using sample test items, additional sample items can be created. Examples of additional procedures include providing sufficient time to establish rapport with shy children, waiting to give test directions until it is obvious the student is focused on the examiner, eliminating extraneous auditory and visual distractions, and using interpreters familiar with a student's primary language and dialect. Unfortunately because of the limited number of nonverbal tests available, and considering the technical adequacy of some measures at certain ages, use of language-reduced measures may be the only or the best option, e.g., preschoolers with language expression problems.

To differentiate between nonverbal tests and tests requiring listening comprehension but not oral expression, in this manuscript, we use the terms *nonverbal* and *language-reduced* as McCallum suggested. In response to the need for nonverbal intellectual assessment, numerous nonverbal tests of cognitive ability and tests of cognitive ability with nonverbal components have been published. Although having numerous options is usually desirable, comparing examiner manuals across the many variables relevant to test selection for these students may be unnecessarily burdensome for practicing school psychologists. Although the information that follows is available in the examiner's manual for each test, the purpose of this manuscript is to present the information in a consolidated document to help school psychologists select appropriate nonverbal or language-reduced tests of cognitive ability, as well as understand the tests' strengths and limitations. Tests were evaluated in terms of standardization samples, psychometric properties, types of directions used, and responses required of students, as well as other test characteristics relevant to meeting student needs that may not be addressed in examiner manuals, including adequacy of floors and item gradients, and percentage of timed items.

Because no single measure should be the basis for conclusions regarding a student's cognitive ability, and because results from a particular measure may be questionable (e.g., student fatigue, a limited number of test items), besides evaluating the seven nonverbal or language-reduced tests, we also provide tables describing eight additional measures with language-reduced components. Whereas these latter tests were not explicitly constructed as nonverbal tests of cognitive ability, use of their language-reduced components can provide supplementary data to increase the sample of student performance obtained.

Method

Nonverbal and language-reduced tests of cognitive ability for students within the age range of birth through 18 years were reviewed. Overall results are described because they provide the best sample of performance. The following criteria were used to evaluate the tests.

The date when normative data were collected is important. Flynn (1984, 1998) demonstrated that if test norms are not updated periodically, examinees receive inflated test results compared with prior generations. This effect is particularly a concern for examinees in the lower ranges of intelligence (Kanaya et al. 2003; Zhou et al. 2010). Salvia et al. (2010) suggested that ability tests more than 15 years old are too old to be representative. Consequently, only tests with normative data collected within the past 15 years were reviewed. Several examiner manuals did not indicate when normative data were collected. However, their copyright dates suggest the data were collected within the past 15 years.

To be representative, demographic data for the standardization sample should be similar to U.S. Census data in terms of geographic distribution, race/ethnicity, gender, urban/rural residence, socioeconomic status (SES; defined as parents' education or occupation or both), percentage of students with impairments, and the number of participants per age level. If students with a cognitive impairment are underrepresented, norms may be inflated. At least 100 participants per age or grade level should be included to guarantee stability, represent infrequent characteristics, and enable the calculation of a full range of derived scores (Salvia et al. 2010). Each measure was evaluated using these criteria.

Psychologists tend to agree (e.g., Sattler 2008; Salvia et al. 2010) that when making important decisions regarding students, the minimum reliability coefficient for acceptable reliability for overall results is .90. We used this criterion to evaluate tests' internal consistency, test–retest, and alternate-form reliability.

Measures also were evaluated for construct, content, and concurrent validity. Methods test authors used to support construct validity are denoted in tables. In the text, independently conducted investigations of the structural validity of measures were noted and described briefly, if available. There are data to suggest contemporary measures of cognitive ability are overfactored (i.e., they purportedly measure more factors than they do) when subjected to rigorous investigations of structural validity (Frazier and Youngstrom 2007), though there are relatively few such investigations of nonverbal measures of cognitive ability, perhaps because most reviewed in the current paper purport to measure a unitary factor. For concurrent validity, the number of other measures of cognitive ability used for comparison is noted, but examiners are encouraged to consult examiner manuals to determine whether the comparison tests are valid.

Inadequate test floors may overestimate performance at some ages. Adequate floors are those where a raw score of 1 results in a standard score two or more standard deviations below the mean (Bracken 1987, 1988). Tests with steep item gradients may over or underestimate students' performance and not discriminate well between those with deficits and those without. Bracken (1987, 1988) suggested an item gradient is too steep when a change of 1 raw score point results in a change greater than 1/3 of a standard deviation in standard or scaled score points. These criteria were used to evaluate adequacy of floors and item gradients. To decrease the probability of over- or underestimating student performance, consideration of such information is necessary.

In addition, the age range of the test, mode for directions, response mode, percent of timed items, and subgroups for whom data were presented are described. Some measures said to be nonverbal use oral directions, but some students who require a nonverbal measure also need nonverbal directions provided through gestures or sign language. The type of student response required is critical. Some measures require pointing, others manipulation of materials, and several include imitation and paper and pencil tasks. For students with motor impairments and communication concerns, some response modes can be problematic. The percentage of timed items is important for some students, e.g., those who are culturally diverse with different conceptions of time, those who are easily distracted, reluctant to participate, respond impulsively, or who have motor impairments. Simeonsson et al. (2001) suggested that tests with timed items be avoided for students who are deaf, because if timed these students may respond as quickly as they can, disregarding accuracy and negatively affecting their performance.

Results

Evaluation summaries are presented for seven nonverbal and language-reduced measures. Following the summaries are tables with specific information for each variable. Table 1 presents descriptions of the tests' standardization samples. Table 2 describes their reliability, and Table 3 presents validity information. Table 4 addresses variables other than technical adequacy that may influence test selection because of the characteristics of particular students. Tables 5, 6, 7, 8 describe the corresponding information for eight tests with language-reduced components.

The Bayley Scales of Infant and Toddler Development—Third Edition (BSID-III; Bayley 2006) is for children from birth through 42 months. The test is well standardized with a large sample for each 1-year age level. The sample is similar to census data except for lack of data for urban/rural residence. Extensive information is provided in support of the test's validity. Adequate floors begin at 16 days and there are no item gradient problems. Data are provided describing how nine subgroups of children performed on the test. The toys and tasks are engaging for young children. Tasks are presented by showing materials or through oral directions. For this language-reduced measure, 86 of the 91 items do not require speech. Children respond by orienting, habituating, manipulating toys, or pointing.

Concerns include low test–retest correlations at all ages and data are provided for age groups rather than for each age level. Twenty-three percent of items are timed which could be problematic for some young children who are, for example, inattentive. No data are presented for children with hearing impairments. Because directions for administration and scoring are complicated, the test requires considerable practice to ensure valid results. Considering the stability of results over time and lack of data on long-term predictive validity, caution is warranted in interpreting results. As for all tests for infants and preschoolers, repeated assessment over time provides a better description of ability than results from a single assessment.

The *Comprehensive Test of Nonverbal Intelligence—Second Edition* (CTONI-2; Hammill et al. 2009) was developed for ages 6 through 89. Strengths include a large representative norm sample similar to census data on all variables but urban/rural residence, for which no data are presented. Reliability is adequate for internal consistency and test–retest for ages 8 through 16. Considerable validity evidence is provided, and floors and item gradients are adequate. The test is easy to use and has no timed items. Examinees point to indicate responses. Oral or signed instructions are recommended, but pantomimed instructions are optional if necessary. Thus, this is intended primarily as a language-reduced test with oral or signed instructions "whenever possible." However, detailed, easy-to-use pantomimed instructions, including pictures, appear in the appendices for use with students who cannot follow oral or signed instructions, making the test a nonverbal measure. Data are presented for seven subgroups of students including those with hearing impairments. One percent of the norm sample involved students with a hearing impairment and data are presented on internal consistency, concurrent validity, and discriminative validity for these students, which is more information than in other tests.

The fact that data on use of the test as a nonverbal measure are limited should be considered when interpreting results. Additional concerns include the lack of test–retest data for ages 6 and 7. Thus, results for these ages should be interpreted cautiously because of lack of information on stability of these results over time. An independent investigation of the structural validity of the CTONI-2 suggested its results should only be interpreted at the level of the overall score as the Pictorial and Geometric dimensions were not supported by exploratory factor analysis (McGill 2016). Why students with hearing impairments score nearly a standard deviation lower than their hearing peers is unclear and should be considered when reporting results for these students.

Table 1 Standardization sample for tests where all items require nonverbal responses

Test	Years of sampling	Geographic distribution	Race/ ethnicity	Gender	SES	Urban/ rural	Students with impairments	*N* per 1-year interval
BSID-III Cognitive Composite	2004	Yes	Yes	Yes	Yes	No data	10 %	100+
CTONI-2 Full-Scale IQ	2007–08	Yes	Yes	Yes	Yes	No data	About 9 %	100+
Leiter-3 Cognitive Battery	Began 2010	Yes	Yes	Yes	Yes	Yes	11 %	Too few at most age levels
PTONI Nonverbal Index	2005–06	Yes	Yes	Yes	Yes	No data	13 %	100+
TONI-4 IQ	Not reported Copyright 2010	Yes	Yes	Yes	Yes	No data	13 %	Ages 6–18 = 100+
UNIT-2 FSIQ	2010–2015	Yes	Yes	Yes	Yes	No data	19 %	<100 at some age levels
WNV Full-Scale IQ	Not reported Copyright 2006	Yes	Yes	Yes	Yes	No data	<4 %	Ages 4–12 = 100+

BSID-III Bayley Scales of Infant Development—Third Ed., *CTONI-2* Comprehensive Test of Nonverbal Intelligence—Second Ed., *Leiter-3* Leiter International Performance Scale—Third Ed., *PTONI* Primary Test of Nonverbal Intelligence, *TONI-4* Test of Nonverbal Intelligence—Fourth Ed., *UNIT-2* Universal Nonverbal Intelligence Test—Second Ed., *WNV* Wechsler Nonverbal Scale of Ability

The *Leiter International Performance Scale—Third Edition* (Leiter-3; Roid et al. 2013) is for examinees ages 3 years through 79 plus. One strength of the test is that the norm sample appears representative on all variables. Reliability data for internal consistency are adequate. Substantial validity evidence is provided, and floors and item gradients are adequate. A strength of the Leiter-3 is that it is one of the very few nonverbal tests available. Instructions are pantomimed and students respond to stimulus pictures by placing blocks into a tray. Tasks and materials are engaging and none of the items are timed. Data are presented for 13 subgroups of students; those with hearing impairments had mean scores similar to those of the norm group.

Concerns for the Leiter-3 include the small number of participants for each age level. Data are presented in 2- or 3-year intervals with the number per interval ranging from 94 to 187. Thus, there were fewer than 100 children at many age levels. Composite test–retest correlations are high, but the retest interval averaged only 7 days and data were collapsed across 3–5 age groups. Considerable familiarity with the instructions is required to administer the test without difficulty. The manual states that examiners should be familiar enough with the test to administer it without using the manual. Although results for students with hearing impairments are similar to those of their hearing peers, and the nonverbal format could be beneficial

Table 2 Reliability for tests where all items require nonverbal responses

Test	Internal consistency	Test–retest (retest interval)
BSID-III Cognitive Composite	$r = .79–.93$	$r = .71$ for ages 2–4 months, $r = .77$ for ages 9–13 months, $r = .86$ for ages 19–26 months and 33–42 months (2–15 days)
CTONI-2 Full-Scale IQ	$r = .92–.96$	$r = .93$ for ages 8–9, $r = .92$ for ages 10–16, $r = .86$ for ages 17–60 (2–4 weeks)
Leiter-3 Cognitive Battery	$r = .70–.96$	$r = .96$ for ages 3–6, $r = .98$ for ages 7–11, $r = .94$ for ages 12–16, $r = .94$ for 17–29 ($M = 7$ days)
PTONI Nonverbal Index	$r = .91–.95$	$r = .95$ for ages 3–4, $r = .98$ for ages 3–6, $r = .93$ for ages 8–9 (2 weeks)
TONI-4 IQ	Form A $r = .94–.97$ Form B $r = .93–.97$	For school age: $r = .88$ Form A, .90 Form B (1 or 2 weeks)
UNIT-2 FSIQ	$r = .97–.99$	$r = .94$ for ages 5–8, $r = .92$ for ages 9–13, $r = .94$ for ages 14–17, $r = .96$ for 18–21 ($M = 17.8$ days)
WNV Full-Scale IQ	$r = .87–.94$	$r = .84$ for ages 4 to 7 (10–31 days), $r = .86$ for ages 8 to 21 (10–22 days)

BSID-III Bayley Scales of Infant Development—Third Ed., *CTONI-2* Comprehensive Test of Nonverbal Intelligence—Second Ed., *Leiter-3* Leiter International Performance Scale—Third Ed., *PTONI* Primary Test of Nonverbal Intelligence, *TONI-4* Test of Nonverbal Intelligence—Fourth Ed., *UNIT-2* Universal Nonverbal Intelligence Test—Second Ed., *WNV* Wechsler Nonverbal Scale of Ability

Table 3 Validity for tests where all items require nonverbal responses

Test	Construct	Content	Concurrent	Adequate floors begin	Item gradient problems
BSID-III Cognitive Composite	Factor analysis results, correlated with language and adaptive behavior, discriminated among atypical groups	Based on prior editions, expert review	Based on prior editions, expert review	16 days	No
CTONI-2 Full-Scale IQ	Scores increased with age; theory-consistent group differences; correlated with achievement	Based on similar tests and relevant theories, conventional item analysis	Correlated with 3 other intelligence tests	6–0	No
Leiter-3 Cognitive Battery	Scores increased to age 15, factor analysis results, discriminated among atypical groups	Based on prior edition, expert review, item-response theory	Correlated with 3 other intelligence tests	3–0	No
PTONI Nonverbal Index	Scores increased with age; theory-consistent group differences; correlated with achievement	Based on similar tests and relevant theories, conventional item analysis	Correlated with 3 other intelligence tests	3–0	No
TONI-4 IQ	Scores increased with age; theory-consistent group differences; factor analysis results, correlated with achievement	Conventional item analysis, factor analysis results	Correlated with 2 other intelligence tests	Form A 7–0 Form B 6–6	No
UNIT-2 FSIQ	Scores increased with age; theory-consistent group differences; factor analysis results, correlated with achievement	Conventional item analysis, differential item functioning analysis	Correlated with 7 other intelligence tests	5–0	No
WNV Full-Scale IQ	Theory-consistent group differences	Based on literature review; adaptations from other intelligence tests	Correlated with 6 other intelligence tests	4–0	Yes for matrices and recognition

BSID-III Bayley Scales of Infant Development—Third Ed., *CTONI-2* Comprehensive Test of Nonverbal Intelligence—Second Ed., *Leiter-3* Leiter International Performance Scale—Third Ed., *PTONI* Primary Test of Nonverbal Intelligence, *TONI-4* Test of Nonverbal Intelligence—Fourth Ed., *UNIT-2* Universal Nonverbal Intelligence Test—Second Ed., *WNV* Wechsler Nonverbal Scale of Ability

for these students, additional technical adequacy data (e.g., stability reliability) for these students would be beneficial.

The *Primary Test of Nonverbal Intelligence* (PTONI; Ehrler and McGhee 2008) was developed for ages 3 through 9. A strength of the test is its adequate sample size that was representative on all demographic variables except data on urban/rural residence are not included. Internal consistency is adequate. Test–retest reliability is strong for all ages except no data are presented for 7-year-olds. Substantial validity evidence is provided. Floors and item gradients are adequate. For this language-reduced test, directions are delivered orally and students point to indicate their responses. The PTONI has no timed items. The test is quick to administer requiring only 5 to 15 min. Data are provided on 11 subgroups of children.

Although quick to administer, a concern with the PTONI is that it is not as comprehensive as some other tests (i.e., provides a limited sample of skills). Also an issue to consider in test selection is that the instruction used for many items is “Find the one that does not belong.” This is an abstract direction that a number of young children or low-functioning children may not understand. Test–retest correlations were high, but data are provided for small age groups rather than each age level and 7-year-olds were not included. Why children with hearing impairments score nearly a standard deviation lower than their hearing peers on this test is unclear.

The *Test of Nonverbal Intelligence—Fourth Edition* (TONI-4; Brown et al. 2010) is for ages 6 through 89. The test has an adequately sized norm sample for school-age students. This is the only test reviewed which has two forms. The norm sample is representative except urban/rural residence is not addressed. Internal consistency correlations are adequate. The test–retest correlation is adequate for Form B. Substantial information is presented regarding validity. Adequate floors begin at 7–0 for Form A and at 6–6 for Form B; there are no problems with item gradients. Instructions can be oral or pantomimed; examinees respond by pointing. One study suggested the oral and pantomimed instructions yield similar results. For 23 % of the sample, norms were collected using pantomimed instructions. Thus, depending on whether the test is given using oral or pantomimed directions, the test is either a language-reduced or

Table 4 Descriptions of tests where all items require nonverbal responses

Test	Age range	Directions	Response required	Percent of timed items	Mean difference for HI	Clinical samples
BSID-III Cognitive Composite	Birth–42 months	Oral and by presenting toys	Orienting and habituating, manipulating materials and pointing	23 %	No data	Down syndrome, PDD, CP, SLD, DD, prematurity, prenatal alcohol exposure, small for gestational age, asphyxiation at birth
CTONI-2 Full-Scale IQ	6–89 years	Oral or pantomime	Pointing	0 %	14 points lower	Gender, race/ethnicity, ADHD, LD, CI
Leiter-3 Cognitive Battery	3–79+	Pantomime	Manipulating materials	0 %	3 points lower	HI, SLD, motor delay, TBI, CI, ADHD, gifted, LD, ELL, autism
PTONI Nonverbal Index	3–9 years	Oral	Pointing	0 %	13 points lower	Gender, race/ethnicity, gifted, alternative language, articulation disorder, ADHD, LD, language disorder, PDD
TONI-4 IQ	6–89 years	Oral or pantomime	Pointing	0 %	No data	Gender, race/ethnicity, ELL, gifted, LD, ADHD, physical impairment, SLD, CI
UNIT-2 FSIQ	5–21 years	Pantomime	Manipulating materials and pointing	8 %	10 points lower	Gifted, ADHD, HI, ASD, LD, LI, ED, ESL, CI
WNV Full-Scale IQ	4–21 years	Pictures, oral if needed	Manipulating materials, drawing, and pointing	Ages 4–7 = 14 %; 8–21 = 18 % plus Coding	No significant difference	Gifted, CI, LD, language disorders, ELL, HI

BSID-III Bayley Scales of Infant Development—Third Ed., *CTONI-2* Comprehensive Test of Nonverbal Intelligence—Second Ed., *Leiter-3* Leiter International Performance Scale—Third Ed., *PTONI* Primary Test of Nonverbal Intelligence, *TONI-4* Test of Nonverbal Intelligence—Fourth Ed., *UNIT-2* Universal Nonverbal Intelligence Test, Second Ed., *WNV* Wechsler Nonverbal Scale of Ability, *AD* Asperger's disorder, *ADHD* attention-deficit/hyperactivity disorder, *CI* cognitive impairment, *CP* cerebral palsy, *DD* developmental delay, *ED* emotional disturbance, *ELL* English language learner, *ESL* English as a second language, *HI* hearing impairment, *LD* learning disabilities, *LI* language impairment, *MI* motor impairment, *PDD* pervasive developmental disorder, *SEI* serious emotional impairment, *SLD* speech and language delay

nonverbal measure. The test has no timed items. The test is easy to administer and requires only about 15 min. No information is presented for performance of students with hearing impairments on the TONI-4, although on a prior version they scored on average 2 points lower than hearing students. Data are presented for 10 other subgroups of students.

The test–retest correlation was low for Form A. Another concern is that all alternate-form reliability correlations were less than .90, suggesting different forms yield somewhat different results. Although the test is quick to administer, it is less comprehensive than most other measures and may best be used for screening or as a supplement to other measures.

The *Universal Nonverbal Intelligence Test—Second Edition* (UNIT-2; Bracken and McCallum 2016) was developed for ages 5 through 21. The norm sample is representative on a number of important variables; however, urban/rural residence is not addressed. Reliability data are strong for both internal consistency and test–retest reliability. Floors and item gradients are adequate. On the UNIT-2, instructions are delivered via standardized gestures and students respond by manipulating materials and pointing.

The UNIT-2 does not report data on the number of students per age level. Considering the total number of students in the norm sample and number of age levels covered by the test, there are at least some age levels with fewer than 100 students. Other concerns include the fact test–retest data are reported by age group. Though substantial validity evidence is provided for the UNIT-2, confirmatory factor analysis results did not consistently support its three factor model (i.e., Reasoning × Memory × Quantitative) at all age levels. In fact, the three factor model had unacceptable fit indexes for the age 5–7-,

Table 5 Standardization sample for tests where components require nonverbal responses

Test	Years of sampling	Geographic distribution	Race/ethnicity	Gender	SES	Urban Rural	Students with impairments	*N* per 1-year Age level
DAS-2 Special Nonverbal Composite	2005	Yes	Yes	Yes	Yes	No data	Included, but descriptive data are not provided	Ages 2–4: 200+; ages 5–17: 200
DTLA-P:3 Verbal Reduced Composite	2002–2003	Yes	Yes	Yes	Yes	No data	8 %	100+
KABC-II Nonverbal Index	2001–2003	Yes	Yes	Yes	Yes	No data	14.5 %	Ages 3–14: 200+;ages 5–18: 100+
M-P-R Developmental Index	Not reported Copyright 2004	No	Yes, but low for African Americans	Yes	Yes	No data	11 %	100+
SB5 FR, QR, VS, and WM subtests in the NV domain	2001–2002	Yes	Yes	Yes	Yes	Not representative	6.8 %	Ages 2–16: 200+; 17–20: 122
WAIS-IV Perceptual Reasoning Index	Not reported Copyright 2008	Yes	Yes	Yes	Yes	No data	Not addressed in manual	Ages 16–17, 18–19: 200
WISC-V Nonverbal Index	2013–2014	Yes	Yes	Yes	Yes	No data	6.5 %	200
WPPSI-IV Nonverbal Index	2010–2012	Yes	Yes	Yes	Yes	No data	No	100+

DAS-2 Differential Ability Scales—Second Ed. (Elliott, 2008), *DTLA-P:3* Detroit Test of Learning Aptitude-Primary—Third Ed. (Hammill & Bryant, 2005), *KABC-II* Kaufman Assessment Battery for Children—Second Ed. (Kaufman & Kaufman, 2004), *M-P-R* Merrill-Palmer-Revised, *SB5* Stanford-Binet—Fifth Ed. (Roid, 2003) (*FR* Fluid Reasoning, *QR* Quantitative Reasoning, *VS* Visual-Spatial Processing, and *WM* Working Memory in the Nonverbal Domain), *WAIS-IV* Wechsler Adult Intelligence Scale—Fourth Ed. (Wechsler, 2008), *WISC-IV* Wechsler Intelligence Test for Children—Fifth Ed. (Wechsler, 2014), *WPPSI-IV* Wechsler Preschool and Primary Scale of Intelligence—Fourth Ed (Wechsler, 2012)

11–13-, and 18–21-year groups. One subtest contains timed items, resulting in a total of 8 % of items. Students with hearing impairments scored about 10 points lower than their hearing peers.

The *Wechsler Nonverbal Scale of Ability* (WNV; Wechsler and Naglieri 2006) was developed for ages 4 through 21. For the norm sample, 100 students were included per age level through age 12; 75 were included for ages 17 and up. The

Table 6 Reliability for tests where components require nonverbal responses

Test	Internal consistency	Test–retest (retest interval)
DAS-2 Special Nonverbal Composite	*r* = .93–.97	*r* = .85 for ages 3–6 through 4, *r* = .87 for ages 5 through 8, *r* = .92 for ages 10 through 11, and ages 14 through 15 (1–9 weeks)
DTLA-P: 3 Verbal Reduced Composite	*r* = .86–.90	*r* = .81 for ages 3–9 (1 week)
KABC-II Nonverbal Index	*r* = .85–.95	*r* = .72 for ages 3–5, *r* = .87 for ages 7–12, *r* = .87 for ages 13–18 (12–56 days)
M-P-R Developmental Index	*r* = .97–.98	*r* = .89 for ages 3–70 months (about 3 weeks)
SB5 FR, QR, VS, and WM in the NV domain	*r* = .72–.93	*r* = .76–.88 (2–5 and 6–20 years: about 5–8 days)
WAIS-IV Perceptual Reasoning Index	*r* = .94–.95	*r* = .86 for ages 16–29 (8–82 days)
WISC-V Nonverbal Index	*r* = .95–.96	*r* = .85 for ages 6–7, *r* = .85 for ages 8–9, *r* = .86 for ages 10–11, *r* = .92 for ages 12–13, *r* = .89 for ages 14–16 (9–82 days)
WPPSI-IV Nonverbal Index	*r* = .94–.96	*r* = .82 for ages 2–6 through 3–11, *r* = .90 for ages 4–0 through 5–5, *r* = .94 for ages 5–6 through 7–7 (7–48 days)

DAS-2 Differential Ability Scales—Second Ed., *DTLA-P:3* Detroit Test of Learning Aptitude-Primary—Third Ed., *KABC-II* Kaufman Assessment Battery for Children—Second Ed., *M-P-R* Merrill-Palmer-Revised, *SB5* Stanford-Binet—Fifth Ed. (*FR* Fluid Reasoning, *QR* Quantitative Reasoning, *VS* Visual-Spatial Processing, and *WM* Working Memory in the Nonverbal Domain), *WAIS-IV* Wechsler Adult Intelligence Scale—Fourth Ed., *WISC-IV* Wechsler Intelligence Test for Children—Fifth Ed., *WPPSI-IV* Wechsler Preschool and Primary Scale of Intelligence—Fourth Ed

Table 7 Validity for tests where components require nonverbal responses

Test	Construct	Content	Concurrent	Adequate floors begin	Item gradient problems
DAS-2 Special Nonverbal Composite	Correlated with achievement	Based on prior edition	Correlated with 3 other intelligence tests	2–6	No
DTLA-P:3 Verbal Reduced Composite	Scores increased with age, prior editions correlated with achievement, discriminated among atypical groups	Review of literature, item analysis	Correlated with prior editions and 10 other intelligence tests	3–9	No
KABC-II Nonverbal Index	Theory-consistent group differences; correlated with achievement	Based on prior edition and cognitive theories	Correlated with 6 other intelligence tests	3–0	Periodic problems for many ages for seven subtests
M-P-R Developmental Index	Scores increased with age, factor analysis results, discriminated among atypical groups	Based on prior edition, expert review, item response theory	Correlated with 3 other intelligence tests	1 month	Yes
SB5: FR, QR, VS, and WM in the NV domain	Theory-consistent group differences; correlated with achievement and adaptive behavior; scores increased with age; factor analysis results	Based on prior editions, expert review, based on previous editions, item analysis	Correlated with 6 other intelligence tests	2–0, except Quantitative Reasoning which is 3–4	Several problems until 11–0 and several for high functioning older students
WAIS-IV Perceptual Reasoning Index	Theory-consistent group differences; correlated with achievement, memory, and neuropsychological status; supported by factor analysis	Based on prior editions, literature review, expert consultation	Correlated with 2 other intelligence tests	16–0	Some problems on Matrix Reasoning and Visual Puzzles for high functioning students
WISC-V Nonverbal Index	Theory-consistent group differences; correlated with achievement, adaptive behavior, and behavioral/emotional functioning; supported by factor analysis	Based on prior editions, literature review, expert consultation, and theoretical rationale	Correlated with 4 other intelligence measures	6–0	A few scattered problems throughout the subtests
WPPSI-IV Nonverbal Index	Theory-consistent group differences, supported by factor analysis	Based on prior editions, literature review, expert consultation, and theoretical rationale	Correlated with 5 other intelligence measures	3–3 (one supp. subtest 4–3)	A few scattered problems throughout the subtests

DAS-2 Differential Ability Scales—Second Ed., *DTLA-P:3* Detroit Test of Learning Aptitude-Primary—Third Ed., *KABC-II* Kaufman Assessment Battery for Children—Second Ed., *M-P-R* Merrill-Palmer-Revised, *SB5* Stanford-Binet—Fifth Ed. (*FR* Fluid Reasoning, *QR* Quantitative Reasoning, *VS* Visual-Spatial Processing, *WM* Working Memory in the Nonverbal Domain), *WAIS-IV* Wechsler Adult Intelligence Scale—Fourth Ed., *WISC-IV* Wechsler Intelligence Test for Children—Fifth Ed., *WPPSI-IV* Wechsler Preschool and Primary Scale of Intelligence—Fourth Ed.

norm sample is similar to census data except students with impairments are underrepresented and no data on urban/rural residence are included. The WNV has considerable evidence in support of validity. The test has adequate floors. This nonverbal test is unique in employing picture sequences to convey directions. Supplemental oral prompts may be used, if needed. Students respond by manipulating materials, drawing, and pointing. The test consists of either a two- or four-subtest option. Data are provided for seven subgroups of students; no significant difference was found between students with hearing impairments and their hearing peers.

One concern about the WNV is that test–retest reliability is low for all ages. Item gradient problems exist on the Matrices and Recognition subtests. Also, depending on the age of the examinee, either 14 or 18 % of items are timed along with the Coding subtest.

Table 8 Description of tests where components require nonverbal responses

Test	Age range	Directions	Response required	Percent of timed items	Mean difference for students with HI	Clinical samples
DAS-II Special Nonverbal Composite	2–6 through 17	Oral or American Sign Language	Manipulating materials, pointing, and drawing	Ages 2–6 through 3–5 45 %; 3–6 through 17:21 %	No significant difference with directions in American Sign Language	Gifted, CI, reading, mathematics, and written language disorders, ADHD, ELL, developmentally at risk, SLD
DTLA-P: 3 Verbal Reduced Composite	3–9 years	Oral	Pointing, drawing, and imitating gestures	0 %	No data	Autism/CI, SLD, LD, ED
KABC-II Nonverbal Index	3–18 years	Oral or pantomime	Manipulating materials, pointing, and imitating gestures	Ages 3–4 = 17 %; 5 = 13 %; 6 = 27 %; 7–18 = 56 %	About 9 points lower	Gender, parental education, race/ethnicity, LD, CI, autistic, ADHD, ED, gifted
M-P-R Developmental Index	1 month through 6–6	Oral or oral with gestures	Manipulating materials and pointing	9 %	15 points lower	CI, prematurity, SLD, MI, autism
SB5 FR, QR, VS, and WM in the NV domain	2–89 years	Oral	Manipulating materials and pointing	10 %	No data	Gifted, CI, DD, autism, ELL, SLD, ADHD, SEI, MI
WAIS-IV Perceptual Reasoning Index	16–‐90 years	Oral	Manipulating materials, pointing, and drawing	61 %	No data	Gifted, CI, LD, ADHD, TBI, autism, AD, ED, dementia
WISC-V Nonverbal Index	6–16 years	Oral	Manipulating materials, pointing, and drawing	57 % plus Coding	No data	Gifted, CI Mild, CI Moderate, BIF, LD Reading, LD Reading and Written Expression, LD, Math, ADHD, DB, TBI, ELL, ASD Language Impairment, ASD without Language Impairment
WPPSI-IV Nonverbal Index	2–6 through 7–7	Oral	Manipulating materials and pointing	Ages 2–6 to 3–0 = 35 %; 4–0 to 7–7 = 16 % plus Bug Search	No data	Gifted, CI, DD, developmental risk factors, preliteracy concerns, ADHD, disruptive, LD, MLD, ELL, autism, AD

DAS-2 Differential Ability Scales—Second Ed., *DTLA-P:3* Detroit Test of Learning Aptitude-Primary—Third Ed., *KABC-II* Kaufman Assessment Battery for Children—Second Ed., *M-P-R* Merrill-Palmer-Revised, *SB5* Stanford-Binet—Fifth Ed. (*FR* Fluid Reasoning, *QR* Quantitative Reasoning, *VS* Visual-Spatial Processing, *WM* Working Memory in the Nonverbal Domain), *WAIS-IV* Wechsler Adult Intelligence Scale—Fourth Ed., *WISC-IV* Wechsler Intelligence Test for Children—Fifth Ed., *WPPSI-IV* Wechsler Preschool and Primary Scale of Intelligence—Fourth Ed., *AD* Asperger's disorder, *ADHD* attention-deficit/hyperactivity disorder, *ASD* autism spectrum disorder, *BIF* borderline intellectual functioning, *CI* cognitive impairment, *DB* disruptive behavior, *DD* developmental delay, *ED* emotional disturbance, *ELD* expressive language disorder, *ELL* English language learner, *HI* hearing impairment, *LD* learning disabilities, *MI* motor impairment, *MLD* mixed receptive-expressive language disorder, *SEI* serious emotional impairment, *SLD* speech and language delay, *TBI* traumatic brain injury

Discussion

Recommending particular nonverbal or language-reduced tests is difficult because important factors in test selection depend upon the characteristics of the student to be tested, e.g., a culturally diverse student with a different conception of time or a student who needs nonverbal directions. On occasion, the need to address such student characteristics may outweigh the need for a test that meets the criteria for technical adequacy. For example, a test with reliability coefficients of at least .90 might be less critical than nonverbal directions for some students. Thus, the various aspects of technical adequacy as well as important student variables require consideration if the most appropriate measures are to be selected for each student. Hopefully, the summary evaluations and corresponding tables will aid school psychologists in efficiently navigating this process.

This review elucidates several issues related to current nonverbal assessment options. One is the need for additional nonverbal tests, where the entire test has nonverbal directions and requires nonverbal responses. Currently, the only options where this is the case are the Leiter-3, TONI-4, UNIT-2, and WNV. For the CTONI-2, oral instructions or signed instructions are recommended, but optional pantomimed instructions can be used if needed. Like most tests, each measure has limitations for use with some students, e.g., the TONI-4 arguably provides only a limited sample of skills and the Leiter-3 has too few students at various age levels in the standardization sample.

If test authors and publishers would include sufficient information in examiner manuals to enable school psychologists to make informed decisions regarding whether a test is appropriate for a particular student this information would be welcome. For example, only the Leiter-3 provides data on its sample's urban/rural residence. Many manuals no longer address this variable. Yet, Roid and Sampers (2004) found significant differences in children's performance based on community size.

To enable school psychologists to determine whether test results for a particular age level are likely to remain relatively stable until a student's planning meeting is held, it would be beneficial if test–retest reliability correlations were reported for each age level or every other age level. To be of use, retest intervals of at least 2 weeks would be helpful. Although a number of examiner manuals report test–retest correlations of at least .90, data are typically averaged across several age groups rather than reported by age level. Further, some tests have retest intervals of less than 2 weeks, which is of limited use in practice.

If mean differences between hearing students and students with a hearing impairment were routinely reported in examiner manuals, this information would assist in test selection and interpretation of results for these students. These data are provided for only about half of the tests. Differences in mean standard scores for these two student groups range from 0 on the WNV to 14 on the CTONI-2. The differences in performance could be a function of how students with a hearing loss are taught, difficulty they have in understanding test directions, actual differences in performance, differences on certain cognitive tasks, or some combination of these and other factors. Research examining why these differences occur is warranted to enhance our understanding of the cognitive development of students with hearing impairments and improve cognitive assessment for these students.

Whereas nonverbal and language-reduced tests may lead to fairer and more valid estimates of cognitive ability for students with communication difficulties or those who do not speak Standard English, we caution school psychologists against their indiscriminant use for these students. It would be a flagrant oversimplification to suggest school psychologists who use nonverbal tests of cognitive ability for these students are meeting their ethical and legal obligation to conduct nondiscriminatory assessments. Our results suggest data describing how ELLs perform compared with a test's standardization sample vary considerably across tests in terms of how subgroups are described, their age ranges, and the size of the samples. Additional data for ELLs would be of considerable assistance. Ortiz and Ochoa (2005) suggested that nonverbal tests are generally preferable for ELLs because of the reduced language demands but noted that the tests themselves do not fully address issues regarding potential linguistic bias or bias due to acculturation. They added that the performance of these students is also affected by how well the student and psychologist interact nonverbally. Consequently, for some students nonverbal tests may be necessary but not sufficient to obtain accurate results. More recent thinking suggests nonverbal tests of cognitive ability should be considered when the student has no or limited oral language and measures are not available to be administered in the student's dominant language (Carvalho et al. 2014). Outside of tests translated to Spanish, there are few non-English options for assessing students' cognitive ability. Carvalho et al. (2014) further point out that nonverbal tests of cognitive ability do not include bilingual students in standardization samples, are potentially confounded by the communication that does occur between the examiner and examinee, and do not measure a full range of abilities thought to comprise current theories of intelligence (e.g., comprehension knowledge). To meet one's ethical and legal obligation in conducting nondiscriminatory assessments of ELLs, school psychologists are guided to recent resources outlining best practices (Carvalho et al. 2014; Ortiz 2014).

Finally, suggesting which nonverbal or language-reduced tests would be most appropriate for students with communication difficulties or students who do not speak standard English is not possible because recommendations would depend upon each student's individual needs. However, following are several recommendations to help ensure an adequate sampling of a student's cognitive skills as well as accurate interpretation of results.

- Because each of the available measures provides only a limited sample of cognitive skills, when possible, use more than one nonverbal measure.
- Nonverbal measures are necessarily more limited than a verbal measure in terms of the cognitive skills assessed; as noted by DeThorne and Schaefer (2004), it may be best to consider overall results a global index of fluid reasoning and/or visual processing.
- Supplement nonverbal measures with data from any of the eight cognitive tests with language-reduced components appropriate for the student.
- When more than one form of a test is available, consider using both forms.
- Interpret cognitive results in light of data from other types of tests or observational procedures for areas such as adaptive behavior, social skills, and academic performance.
- Interpret cognitive results noting concerns mentioned in this review about the particular tests administered, e.g., students with hearing impairments tend to score on average 10 points lower on this test than their hearing peers.
- Interpret cognitive results considering prior assessment results. For initial assessments, mention that typically repeated assessments over time provide a better sample of a student's performance than a single assessment.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Funding The authors received no financial support for the research, authorship, and/or publication of this article.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Bayley, N. (2006). *Bayley Scales of Infant and Toddler Development* (Third ed.). San Antonio: Psychological Corporation.

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment, 4*, 313–326.

Bracken, B. A. (1988). Ten psychometric reasons why similar tests produce dissimilar results. *Journal of School Psychology, 26*, 155–166.

Bracken, B. A., & McCallum, R. S. (1998) *Universal Nonverbal Intelligence Test*. Itasca, IL: Riverside.

Bracken, B. A., & McCallum, R. S. (2016). *Universal Nonverbal Intelligence Test* (Second ed.). Austin: PRO-ED.

Brown, L., Sherbenou, R. J., & Johnsen, S. K. (2010). *Test of Nonverbal Intelligence* (Fourth ed.). Austin: PRO-ED.

Carvalho, C., Dennison, A., & Estrella, I. (2014). Best practices in the assessment of English language learners. In P. L. Harrison & A. Thomas (Eds.), *Best practices in school psychology VI* (pp. 75–87). Bethesda: NASP Publications.

DeThorne, L. S., & Schaefer, B. A. (2004). A guide to child nonverbal IQ measures. *Journal of Speech-Language Pathology, 13,* 275–290. doi:10.1044/1058-0360(2004/029).

Ehrler, D. J., & McGhee, R. L. (2008). *Primary Test of Nonverbal Intelligence*. Austin: PRO-ED.

Elliott, C. D. (2008). *Differential Abilities Scale* (Second ed.). San Antonio: Psychological Corporation.

Flanagan, D. P., Alfonso, V. C., & Dynda, A. M. (2008). Best practices in cognitive assessment. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology V* (pp. 633–650). Bethesda: NASP Publications.

Flynn, J. R. (1984). The mean IQ of Americans: massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29–51. doi:10.1037/0033-2909.95.1.29.

Flynn, J. R. (1998). WAIS-III and WISC-III: IQ gains in the United States from 1972 to 1995: how to compensate for obsolete norms. *Perceptual and Motor Skills, 86*, 1231–1239. doi:10.2466/pms.1998.86.3c.1231.

Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: are we overfactoring? *Intelligence, 35*, 169–182. doi:10.1016/j.intell.2006.07.002.

Hammill, D. D., & Bryant, B. (2005). *Detroit Test of Learning Aptitude-Primary* (Third ed.). Austin: PRO-ED.

Hammill, D. D., Pearson, N., & Wiederholt, L. J. (2009). *Comprehensive Test of Nonverbal Intelligence* (Second ed.). Austin: PRO-ED.

Kanaya, T., Scullin, M. J., & Ceci, S. J. (2003). The Flynn Effect and US policies: the impact of rising IQ scores on American society via mental retardation diagnoses. *American Psychologist, 58*, 778–790. doi:10.1037/0003-066X.58.10.778.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children* (Second ed.). Circle Pines: AGS.

McCallum, R. S. (2003). Context for nonverbal assessment of intelligence and related abilities. In R. S. McCallum (Ed.), *Handbook of nonverbal assessment* (pp. 3–21). New York: Kluwer Academic/Plenum Publishers.

McGill, R. J. (2016). Investigation of the factor structure of the Comprehensive Test of Nonverbal Intelligence-Second Edition (CTONI-2) using exploratory factor analysis. *Journal of Psychoeducational Assessment, 34*, 339–350. doi:10.1177/0734282915610717.

Naglieri, J. A., & Otero, T. M. (2012). The Wechsler Nonverbal Scale of Ability. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: theories, tests, and issues* (pp. 436–455). New York: Guilford.

Ortiz, S. O. (2014). Best practices in nondiscriminatory assessment. In P. L. Harrison & A. Thomas (Eds.), *Best practices in school psychology VI* (pp. 61–74). Bethesda: NASP Publications.

Ortiz, S. O., & Ochoa, S. H. (2005). Cognitive assessment of culturally and linguistically diverse individuals. In R. L. Rhodes, S. H. Ochoa, & S. O. Ortiz (Eds.), *Assessing culturally and linguistically diverse students* (pp. 168–201). New York: Guilford.

Reynolds, C. R., Livingston, R. A., & Wilson, V. L. (2006). *Measurement and assessment in the classroom*. Boston: Allyn & Bacon.

Roid, G. H., Miller, L. J., Pomplun, M., & Koch, C. (2013). *Leiter International Performance Scale* (Third ed.). Wood Dale, IL: Stoelting.

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales* (Fifth ed.). Itasca: Riverside Publishing.

Roid, G. H., & Sampers, J. L. (2004). *Merrill-Palmer-revised: scales of development*. Wood Dale: Stoelting.

Salvia, J., Ysseldyke, J., & Bolt, S. (2010). *Assessment in special and inclusive education* (11th ed.). Boston: Houghton Mifflin.

Sattler, J. M. (2008). *Assessment of children: cognitive foundations* (5th ed.). La Mesa: Jerome M. Sattler Publisher.

Simeonsson, R. J., Wax, T. M., & White, K. (2001). Assessment of children who are deaf or hard of hearing. In R. J. Simeonsson & S. L. Rosenthal (Eds.), *Psychological and developmental assessment: children with disabilities and chronic conditions* (pp. 248–266). New York: Guilford.

Wechsler, D. (2012). *Wechsler Preschool and Primary Scale of Intelligence* (Fourth ed.). San Antonio: Harcourt Assessment.

Wechsler, D. (2014). *Wechsler Intelligence Scale for Children* (Fifth ed.). Bloomington: NCS Pearson, Inc..

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale* (Fourth ed.). San Antonio: NCS Pearson, Inc..

Wechsler, D., & Naglieri, J. A. (2006). *Wechsler Nonverbal Scale of Ability*. San Antonio: Harcourt Assessment.

Zhou, X., Zhu, J., & Weiss, L. G. (2010). Peeking inside the "black box" of the Flynn effect: evidence from three Wechsler instruments. *Journal of Psychoeducational Assessment, 28*, 399–411. doi:10.1177/073428291-37340.

Daniel D. Drevon PhD, is an Assistant Professor in the Department of Psychology at Central Michigan University. His research focuses on behavior analytic academic and behavior interventions for school-age children and research synthesis/meta-analysis.

Rachel M. Knight PhD, is an Assistant Professor in the Department of Pediatrics with the University of Michigan Medical School. Her research focuses on behavioral interventions for pediatric feeding disorders and sleep disorders.

Sharon Bradley-Johnson EdD, is a Professor Emeritus in the Department of Psychology at Central Michigan University. Her research interests include psychoeducational assessment, low-incidence disabilities, and cognitive development for infants, toddlers and preschoolers.