Article

Comparing Age- and Grade-Based Norms on the Woodcock–Johnson III Normative Update

Educational and Psychological Measurement
2019, Vol. 79(5) 855–873

DOI: 10.1177/0013164419834607
journals.sagepub.com/home/epm
SAGE

Allyson G. Harrison[1], Kaitlyn Butt[1] and Irene Armstrong[1]

Abstract

There has been a marked increase in accommodation requests from students with disabilities at both the postsecondary education level and on high-stakes examinations. As such, accurate identification and quantification of normative impairment is essential for equitable provision of accommodations. Considerable diversity currently exists in methods used to diagnose learning disabilities, including whether an impairment is normative or relative. This study investigated the impact on impairment classification if grade-based norms were used to interpret identical raw scores compared with age-based norms. Fourteen raw scores distributed equally across the adult range of the Woodcock–Johnson III Normative Update subtests were scored using norms for either age (18-29 years) or grade (13-17). The results indicate that raw scores receive a significantly lower standardized score (and thus percentile ranking) when grade-based norms are used. Furthermore, the difference between age- and grade-based scores increases dramatically as raw scores decrease, and there is a significant interaction between age and grade in the standard scores obtained. This study provides evidence to suggest that using different norms may result in different decisions about diagnoses and appropriate accommodations.

Keywords

assessment, adults, disability, learning disability, normative sample, diagnosis

[1]Queen's University, Kingston, Ontario, Canada

Corresponding Author:
Allyson G. Harrison, Queen's University, 68 University Avenue, Kingston, Ontario, K7L 3N6, Canada.
Email: harrisna@queensu.ca

Over the past 10 years, there has been a marked increase in the number of students diagnosed with disabilities who are seeking academic accommodations in both postsecondary education (Cox, Herner, Demczyk, & Nieberding, 2006; Deloitte, 2017; Harrison & Wolforth, 2012; Ranseen & Parks, 2005; Ready, Chaudhry, Schatz, & Strazzullo, 2012) and on high-stakes examinations (Lauth, Sweeney, & Reese, 2017; Lewandowski, Codding, Kleinmann, & Tucker, 2003; Lovett, 2010; Yellin, 2016). In fact, between 11% and 15% of all students in higher education are reported to have a disability that requires accommodation (Kimball, Wells, Ostiguy, Manly, & Lauterbach, 2016). The most commonly requested accommodation is extra time (Bolt & Thurlow, 2004; Harrison & Wolforth, 2012; Julian, 2005; Keiser, 1998; Kettler, 2012; Lewandowski et al., 2003; Ofiesh, Hughes, & Scott, 2004; Stretch & Osborne, 2005), typically requested due to reported problems with speed of reading or processing information. To obtain such accommodations as an adult, most institutions or testing agencies require assessment data to be recent (i.e., completed in the past 3-5 years) (Gyenes & Siegel, 2014; Lindstrom & Lindstrom, 2017).

Different postsecondary institutions and testing agencies all have differing requirements for disability documentation and differing methods for determining whether accommodations will be granted (e.g., Gyenes & Siegel, 2014; Lindstrom & Lindstrom, 2017; Madaus, Banerjee, & Hamblet, 2010). Due to what has been termed a *documentation disconnect*, students previously provided with accommodations in high school due to a nonvisible disability may not possess the specific documentation required for provision of accommodations at their chosen college or when taking high-stakes examinations, thus requiring reevaluation as an adult (e.g., Gyenes & Siegel, 2014; Madaus et al., 2010).

Individuals with disabilities are entitled to receive reasonable academic accommodations anytime that the functional impairments related to their diagnosed condition interfere with their equal participation in an academic setting. In theory, academic testing accommodations should remove construct-irrelevant variables that interfere with accurate evaluation of knowledge or skills (Fuchs & Fuchs, 2001; Lai & Berkeley, 2012; Roberts, 2012; Thurlow, Thompson, & Lazarus, 2006). When used properly, academic accommodations lead to more valid test scores for examinees with disabilities; however, if given inappropriately, they may give examinees an unfair advantage, allowing them to access significantly more of the test material than their nondisabled peers (Lerner, 2004; Lewandowski, Cohen, & Lovett, 2013; Lovett, 2010).

Accurate identification and quantification of cognitive impairment is therefore important in determining the need for extra time accommodations. Postsecondary disability services offices and testing agencies, for example, seek confirmation that the individual is impaired in some relevant cognitive process such as speed of visual processing, reading ability, or academic fluency. Clinicians typically document impairment by means of scores from standardized tests and identify where the individual is performing below expected levels (Lovett, Gordon, & Lewandowski, 2009); however, comprehensive, psychometrically sophisticated, and agreed-upon guidelines for

identifying and quantifying cognitive impairment do not exist (Iverson & Brooks, 2011). Specifically, it is not clear at what point a score becomes low enough to constitute a disability, how many scores need to be below average in order to confirm a disability, and against what population standard this decision should be made.

A review of recent literature reveals that there is considerable diversity in the methods used to identify those who qualify as disabled (Harrison, 2017). Specifically, there is discord in the legal and educational systems with respect to which diagnostic criteria and appropriate reference group should be used when determining impairment in postsecondary or graduate-level adults. The prevailing conceptualization of impairment, at least at the postsecondary level, is that it must be determined relative to most other people in the general population (i.e., a normative weakness; Lovett et al., 2009). The Americans with Disabilities Act, as amended in 2008 (ADAAA, 2008) uses this normative standard for determining the presence of a disabling condition, noting that ''an individual must have an impairment that prevents or severely restricts the individual from doing activities that are of central importance to most people's daily lives.'' The current *Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition (*DSM-5*; American Psychiatric Association, 2013) notes that for the diagnosis of a specific learning disorder (SLD) an individual demonstrates impairment in some academic ability, demonstrated by academic achievement scores that fall in the lowest 7% to 16% of the general population (American Psychiatric Association, 2013). Similarly, *DSM-5* defines mild neurocognitive disorder as a condition in which an individual performs below the 16th percentile in one or more cognitive domains. *DSM-5* does not, however, specify whether this refers to performance on a single test or a psychometrically consistent cluster of tests that yield an overall index score.

Not everyone, however, agrees with this method of determining disability status (Thomas, 2000). Some professionals, for example, have argued that individuals may be diagnosed as disabled even if they do not demonstrate normative deficits in achievement (e.g., Barber & Mueller, 2011; Gilman et al., 2013; McCallum et al., 2013; McCoach, Kehle, Bray, & Siegle, 2001). From this perspective, impairment is determined relative to the individual's theoretical potential rather than relative to their same-aged peers. Similarly, the former edition of the *DSM* (4th edition, text revision; *DSM-IV-TR*; American Psychiatric Association, 2000) simply required between a 1 to 2 standard deviation discrepancy (plus or minus the standard error of measurement) between the individual's intellectual ability and performance in some area of academic achievement to make a diagnosis of SLD. Despite the fact that use of such discrepancy formulae for determination of disability has been discredited in the literature (see Harrison & Holmes, 2012, for an extended discussion), many clinicians continue to employ this diagnostic method (Gordon, Lewandowski, Murphy, & Dempsey, 2002; Harrison, Lovett, & Gordon, 2013; Weiss, Erikson, & Till, 2017).

Looking specifically at the discrepancy method of diagnosing SLD, Maddocks (2018) showed that many individuals in the normative sample of a widely used achievement test would qualify for the diagnosis of SLD and that individuals with

well above-average intelligence often have areas of strengths and weaknesses in major areas of academic functioning and psychological processing. Indeed, when examining data from the Woodcock–Johnson Tests, Third Edition (WJ-III; Woodcock, McGrew, & Mather, 2001a, 2001b) normative sample, Maddocks (2018) found that just over half (56%) of all cases in the normative data set would be classified as SLD if some type of discrepancy criteria were employed as a diagnostic method. Furthermore, she found that 61.2% of the general normative sample scored below the 25th percentile on at least one academic cluster, raising concerns about simply using one low achievement score as an indication of SLD. Finally, she demonstrated that a very high proportion of otherwise nondisabled individuals with a Full Scale IQ (FSIQ) ≥ 130 would be diagnosed as both ''gifted'' and ''learning disabled'' (gifted LD: a nonspecific term meaning that the individual is both intellectually gifted and also ''underperforming'' in some area of achievement) if using ability–achievement discrepancy criteria. For example, McCallum et al. (2013) suggest a method of identifying ''gifted LD,'' where any observed academic cluster–scaled score need only be lower than predicted based on the standard error of estimate (accounting for correlations between cognitive and academic scores), so that, for example, an FSIQ of 143 and Basic Reading score of 122 would be considered evidence of someone being ''gifted LD.'' In fact, when Maddocks applied the discrepancy criteria suggested by McCallum et al., she found that 89% of the high IQ normative sample of the WJ-III would be identified as ''gifted LD.''

The disagreement as to the referent population also results in different diagnostic outcomes. One correlational study by Giovingo, Proctor, and Prevatt (2005) used scores from 155 American college students on Woodcock–Johnson Tests of Achievement–Third Edition (WJ-III ACH; Woodcock et al., 2001a) and the Woodcock–Johnson Tests of Cognitive Abilities–Third Edition (WJ-III COG; Woodcock et al., 2001b) to compare three diagnostic models of LD using both age- and grade-based norms: (1) intraindividual discrepancy between cognitive and achievement cluster scores and the average of the remaining cognitive and achievement scores on the WJ-III COG and WJ-III ACH; (2) ability–achievement discrepancy between one's General Intellectual Ability score on the WJ-III COG and each of the four broad achievement scores; and (3) underachievement, classified as any WJ-III ACH broad score that was equal to or below the 16th percentile. The results indicated that for both the intraindividual discrepancy and the underachievement models, the grade-based norms yielded significantly more LD diagnoses than their equivalent age-based models.

A later study conducted by Cressman and Liljequist (2014) provides further evidence in support of this finding. In this case, the records of 202 college students who had requested an evaluation of learning disorders were gathered and discrepancy scores were calculated between the Wechsler Adult Intelligence Scale–Third Edition (WAIS-III; Psychological Corporation, 1997) FSIQ and the WJ-III Total Achievement, Broad Reading, Math, and Written Language scores. Overall, use of grade-based norms resulted in significantly lower scores compared with use of

age-based norms. Using the strict *DSM-IV-TR* diagnostic criteria for SLD (e.g., a discrepancy score greater than or equal to 2 standard deviations plus or minus the standard error below the individual's IQ), it was found that the proportion meeting the discrepancy criterion was 7.9% using age norms and 37.6% using grade norms. Furthermore, employing the absolute score method for diagnosis (i.e., the 16th percentile or lower) also resulted in a significantly higher proportion of students identified as impaired if grade-based norms were used to interpret identical raw scores.

Psychometric problems exist when employing grade-based norms to assess performance of individuals who are enrolled in or have graduated from a postsecondary program. This is because less than half of the total U.S. population attends higher education and even fewer (20.8%) graduate from a 4-year institution (U.S. Census Bureau, 2016). This means that the normative sample for postsecondary grade-level students does not represent all individuals in the general population but a sample of individuals whose scores are not normally distributed; their scores typically fall in the upper half of the Gaussian normal distribution. Brooks, Sherman, Iverson, Slick, and Strauss (2011) reviewed the significant problems that can arise when clinicians employ tests with such skewed distributions (i.e., with normative data obtained mainly or exclusively from only one-half of the theoretical normal distribution), especially when attempting to identify impairment relative to most other individuals in the general population. In particular, they note that tests with a highly educated normative sample that exclude a proportion of people falling in the lower half of the normal distribution will consistently identify normally achieving individuals as being impaired. In essence, it is similar to evaluating a person's golfing skills by comparing her not to all female golfers but to only those few who played collegiate-level golf. In such an example, an otherwise average golfer would now appear deficient if compared only with such an elite group. Brooks et al. (2011) show that when low-functioning individuals are excluded from a normative sample, the resulting low end of the new distribution (or lowest percentiles) are then occupied by persons who would have populated the higher percentiles in the normal distribution of all individuals in the general population. Brooks et al. warn that this, in turn, can lead to normally functioning individuals being falsely identified as low functioning or impaired. The question, then, is to what extent might grade-based norms identify otherwise average individuals as impaired?

One popular educational achievement test that allows for test score interpretation using either age- or grade-based norms is the WJ. The WJ-III Normative Update (WJ-III NU; Woodcock, McGrew, Schrank, & Mather, 2007) and the newer WJ-IV (Schrank, Mather, & McGrew, 2014) allow clinicians to interpret scores obtained by young adults based on either age-based norms or norms collected from students in Grades 13 through 17. While the sample sizes for those aged 19 to 29 years and for Grades 13 to 17 are relatively similar (approximately 1,000 people in this age range or higher education samples, respectively), Gregg, Coleman, Lindstrom, and Lee (2007) note that the total number of students included in these norms represents a very small (less than 0.002%) proportion of individuals occupying that subpopulation

in the United States. Additionally, the majority of students representing the college grade distribution were selected from 4-year college or university programs ($N = 976$ for WJ-III, 572 for WJ-IV), while only a small number were selected from 2-year college programs ($N = 186$ for WJ-III, 167 for WJ-IV). In other words, most of the postsecondary norms were collected from a group of individuals who represent only a small, higher functioning proportion of the total population.

The two previous studies that investigated the difference between age and grade norms employed clinical samples, who might not be representative of all young adults attending or having graduated from a postsecondary institution. The present study, therefore, investigated the difference between scores on the WJ-III NU Tests of Achievement when identical raw scores were converted by the WJ-III NU computer scoring program using either age- or grade-based norms. Specifically, we were interested not only in the standard score (and, by association, the percentile rank) changes that might occur when using age- or grade-based norms but also whether differences in obtained scores changed across the spectrum of possible raw scores or across age or grade levels. Based on results from previous studies, we hypothesized that scores obtained by using postsecondary grade-based norms would result in significantly lower scaled (and thus percentile) scores than when age-based norms were employed to interpret the meaning of the same raw scores. Furthermore, we anticipated that this difference would grow larger for scores at the lower levels of proficiency.

In addition, we wondered to what extent test ceilings (i.e., the highest score possible for any age or grade) might artificially create discrepancies that could be interpreted as indications of a young adult being both gifted and SLD, especially if grade-based scores were employed when determining academic achievement.

Finally, we hypothesized that score differences obtained by use of grade versus age norms would be clinically meaningful (i.e., achievement classification categories would change) for a substantial number of individuals and that this could lead to different diagnostic decisions for a sizable number of students.

Method

Materials

Two speed-related subtests and two language processing subtests of the WJ-III NU were used in this investigation. The two speeded tests were as follows:

The Reading Fluency subtest consists of 98 simple sentences primarily describing common animals and objects. Subjects are required to silently read as many of the sentences as possible in 3 minutes, circling Y for ''yes'' or N for ''no'' after each sentence, depending on whether the statement was true or false. All subjects begin at Question 1. *The Math Fluency* subtest requires the subject to solve a series of simple addition, subtraction, and multiplication problems (e.g., $3 + 1 = ?$; $2 \times 0 = ?$) and write their answer down on paper within 3 minutes. All subjects begin at Question 1. According to the technical manual (Woodcock et al., 2007), WJ III NU 1-day test–retest stability scores for these speeded subtests are lower than those for other

nonspeeded tests in this battery. Due to the timed nature of these tests, internal reliability statistics are not provided. No data are provided regarding 1-day test–retest stability for individuals aged 18 to 25 years; however, stability in teenagers and in those older than 26 years ranges from a low of .80 for Reading Fluency in teenagers to a high of .96 for Math Fluency in those older than 26 years.

The two subtests that constitute the Basic Reading Cluster score are Letter-Word Identification and Word Attack. *Letter-Word Identification* is an oral test of reading skills. The subject must correctly pronounce words aloud from an increasingly difficult vocabulary list. *Word Attack* is also a test of oral reading, but it requires the subject to decode and correctly pronounce an increasingly difficult series of nonsense words (e.g., plurp, fronkett) to test phonetic Word Attack skills. According to the technical manual (Woodcock et al., 2007), the internal consistency coefficient for 19-year-old subjects is .90 for Letter-Word Identification (.91 for young adults aged 20 to 29) and .87 for Word Attack (.83 for aged 20 to 29)

Procedure

The WJ-III NU cannot be scored by hand; the Compuscore program for the WJ-III NU, however, allows a clinician to enter unlimited raw scores to calculate standard scores relative to either age- or grade-based norms. As such, 14 hypothetical raw scores were created for *each of the four* subtests of the WJ-III NU most typically relied on when clinical assessment reports recommend extra-time accommodations on postsecondary and high-stakes examinations (i.e., Reading Fluency, Math Fluency, Letter-Word Identification, and Word Attack). (We chose 14 scores per subtest because some of the chosen subtests [i.e., Word Attack] allowed for only 14 raw scores between the basal and ceiling level for all adults.) These 14 raw scores were distributed equally and spanned the entire ''adult'' range of scores for each of these WJ-III NU subtests. That is, the raw scores were spaced equally from the lowest to the highest possible scores for *each of the four* subtests. We assumed only that Case 1 (lowest) would receive credit for all items up to the adult baseline question for each subtest (if applicable) or would be assigned a raw score equivalent to a Grade 4 level performance on that subtest (as it was deemed unlikely that a college student would obtain a score of zero on any of these tests). Of note, standard scores from Letter-Word Identification and Word Attack combine in the computer scoring to create the Basic Reading Cluster composite score.

We therefore calculated the scores as follows: Case 1 obtained the lowest raw score on each of the four subtest (assuming baseline items were passed), and the raw scores for each hypothetical case on each subtest then increased equally for each subsequent case until Case 14 where the highest raw scores were obtained on each of the four subtests. Next, the calculated raw scores for the 14 cases for each subtest were entered manually into four separate Microsoft Excel spreadsheets (one for each subtest), identifying subtest name and raw scores from the lowest (Case 1) to the highest (Case 14). A ''dummy'' subject then had to be created in the WJ-III NU

computer scoring program (Scoring Assistant software) for each proficiency level (from the lowest to the highest raw scores for each of the four WJ-III NU subtests). We assumed that each ''dummy'' subject was male (standard scores for the test are not classified based on gender) and gave all 14 ''dummy'' subjects the same birthdate. Subtest and Cluster scores for each ''dummy'' subject were then calculated using either age- or grade-based norms for each of the 14 hypothetical cases (e.g., ''dummy'' Subject Number 1 has the lowest score on each of the four subtests, ''dummy'' Subject Number 2 has the second lowest, etc.). Finally, we altered each ''dummy'' subject in the computer scoring program to correspond to one of eight possible ages (e.g., one complete set of 14 raw scores for each dummy subjects was scored using norms for ages 18-21, 23, 25, 27, and 29) or five possible grades (Grade 13, Grade 14, Grade 15, Grade 16, and Grade 17 (graduate student level)) to investigate the score differences across various young-adult age-groups over the spectrum of possible age- or grade-level scores.

To ensure interscorer reliability, the first author then reentered the same raw scores into a separate scoring assistant program and independently calculated the subtest and cluster scores for each of eight possible ages and five possible grade levels. Interrater agreement was perfect.

Results

Differences Between Age- and Grade-Based Scores for Each Subtest/ Academic Cluster

For each raw value, a standard score was calculated using both grade- and age-based normative data. Analyses were performed on the difference between age-based standard scores and grade-based ones. Analyses are limited when comparing the difference between age- and grade-based standard scores, as there was only one difference score for each age or grade across ability level (e.g., proficiency). Standard scores have a mean of 100 and a standard deviation of 15. The difference in standard scores between age-based and grade-based norms was always greater than zero; in every case, grade-based norms provided lower standard scores than did age-based norms.

Calculating the difference between age-based and grade-based norms for each possible age-grade pairing resulted in a single difference score between the following pairs: Age 18 and Grade 13, Age 19 and Grade 14, Age 20 and Grade 15, Age 21 and Grade 16, and Age 23 and Grade 17. Assuming that graduate- or postgraduate-level students could be compared with either age or Grade 17 norms (highest grade norms available), we calculated scores for ages 23 to 29 using either Grade 17–based norms or the respective age-based norms (see below).

To examine changes across the paired age/grade difference up to Age 23, a one-way analysis of variance was performed with the difference score as the dependent measure and age/grade as the independent measure, collapsed across the 14 levels of proficiency (see Figure 1). Across all subtests, there is a larger age–grade difference as age and grade increase: Reading Fluency $F(13, 56) = 15.36$, $p < .001$; Math

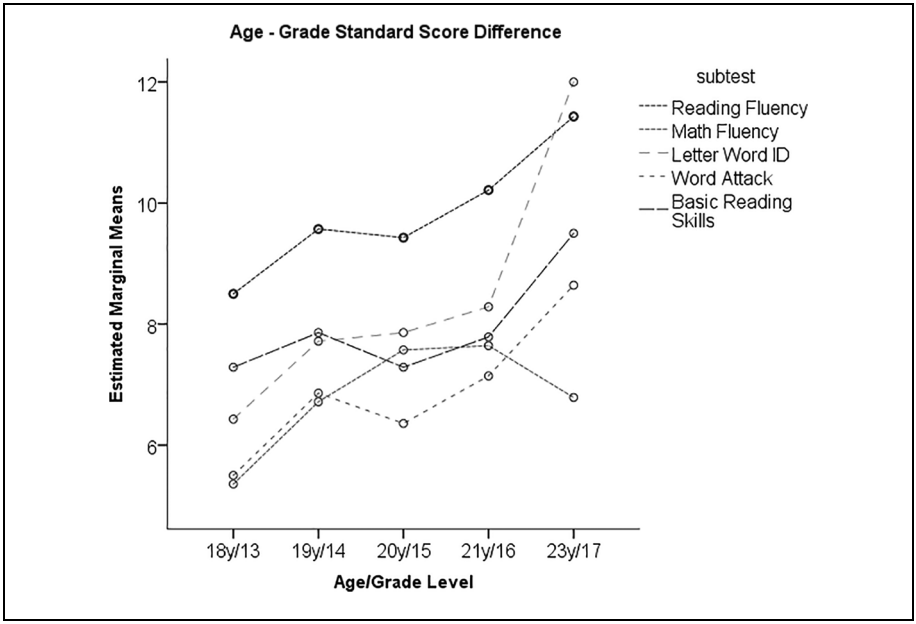

Figure 1. Age minus grade standard score difference for five subtests of the Woodcock–Johnson III Normative Update (WJ-III NU) as a function of age/grade level.

Fluency $F(13, 56) = 3.25$, $p = .001$; Letter-Word Identification $F(13, 56) = 3.085$, $p = .002$; Word Attack $F(13, 56) = 11.14$, $p < .001$; Basic Reading Skills $F(13, 56) = 4.60$, $p < .001$.

We also examined the same groups (ages 18 to 23 years matched to Grades 13 to 17) as a function of proficiency. The dependent measure again is the difference score and the independent measure is proficiency. As shown in Figure 2, we find that with the exception of Math Fluency, each subtest shows significantly larger difference scores at weaker proficiencies than at stronger proficiencies: Reading Fluency $F(4, 65) = 2.56$, $p = .047$; Math Fluency $F(4, 65) = 7.08$, $p < .001$; Letter-Word Identification $F(4, 65) = 13.97$, $p < .001$; Word Attack $F(4, 65) = 3.33$, $p = .015$; Basic Reading Skills $F(4, 65) = 7.04$, $p < .001$.

We then examined changes in standard scores for individuals considered to be in graduate-level or postgraduate-level grades (from ages 23 to 29 years). All standard scores derived from age-based norms were compared with standard scores derived from Grade 17–based norms using the same raw data. The dependent measure is the difference in standard score between age-based and Grade 17–based norms. We found a main effect for age ($p < .001$) such that the difference score is largest for the oldest age (29), and this effect is monotonic. There is also an interaction between age and proficiency level ($p < .001$) such that those with lower raw scores and those who are older show a greater difference than those with higher scores or younger ages.

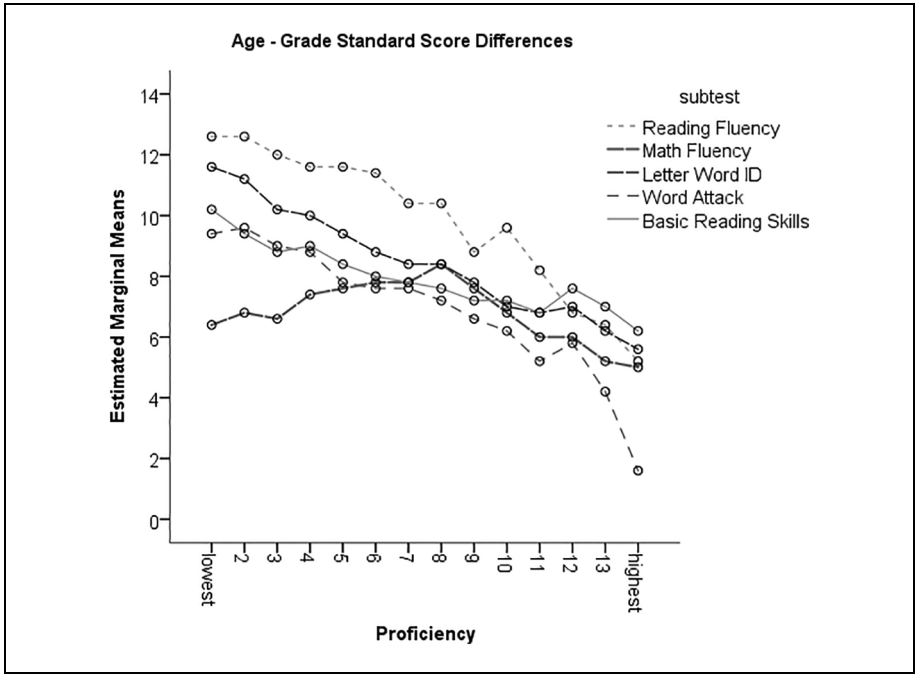

Figure 2. Age minus grade standard score difference for five subtests of the Woodcock–Johnson III Normative Update (WJ-III NU) as a function of proficiency level from worst to perfect performance.

The analyses below (Figure 3) collapsed scores across the subtests. Figure 4 shows the effects of grade versus age by subtest. Note that no statistical analysis has been done on these ''By Subtest'' graphs because there is only one data point contributing to each marker on each graph.

Effect of Test Ceilings on Discrepancy Scores

Given the question of possible classification as gifted and SLD, it was of interest to determine to what extent subtest ceilings might artificially create significant IQ–achievement score differences. Table 1 shows the highest possible scores that can be achieved for each subtest based on age- or grade-based norms. As may be seen, even using age-based norms there are notable test ceilings, and the maximum possible score declines as both age and grade increase.

Classification Change

Finally, we examined how achievement classification might change depending on the norms used (see Table 2). Using standard psychometric classification nomenclature

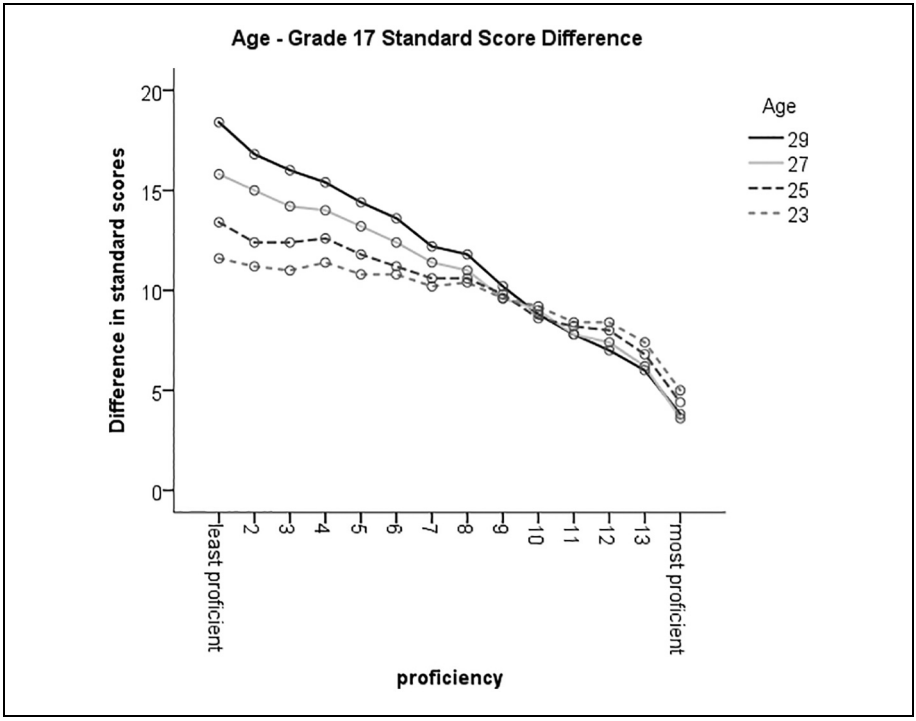

Figure 3. Age minus Grade 17 standard score differences for ages 23, 25, 27, and 29 years versus Grade 17 as a function of level of proficiency from the worst to the perfect scores on the Woodcock–Johnson III Normative Update (WJ-III NU).

(e.g., Mitrushina, Boone, Razani, & D'Elia, 2005), we classified standard scores of 121 or greater as superior, from 80 to 120 as average, and scores below 80 as borderline. As can be seen, many more cases are found to be borderline using grade-based norms. Indeed, the number of raw scores that receive a standard score below 80 is higher using grade-based norms (see Table 2). Using the Letter-Word Identification subtest as an example, no fewer than 9 of the 14 scores we calculated received a standard score below 80 using grade-based norms, whereas fewer than 6 of the 14 scores received a standard score below 80 when age-based norms are used. This observation was true across all five subtests that were analyzed in this study. Of note, almost no scores were classified in the superior range despite our use of the highest scores possible for every subtest.

Discussion

The goal of the present study was to investigate the the difference between obtained standard scores on the WJ-III NU Tests of Achievement when identical raw scores

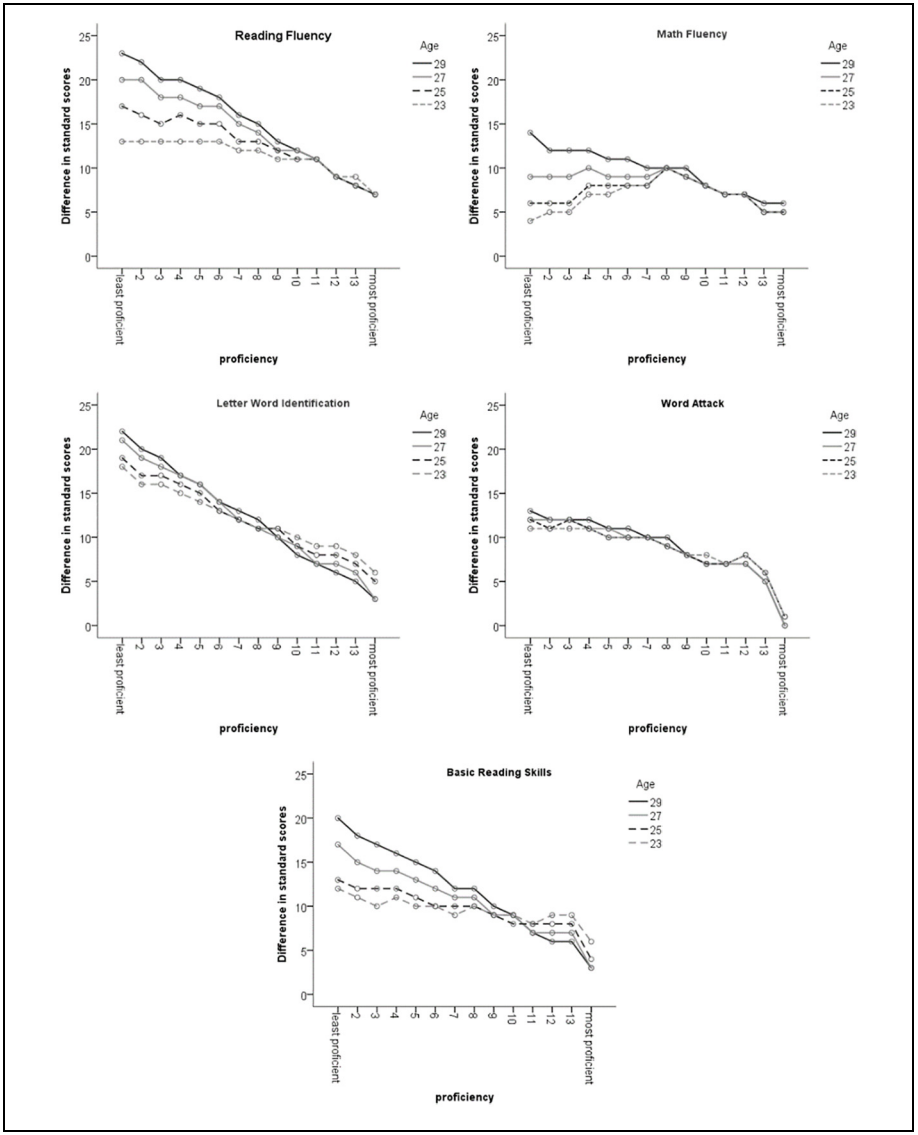

Figure 4. Age minus grade standard score differences for age-based minus Grade 17–based norms as a function of level of proficiency from the worst to the perfect scores shown separately for five subtests of the Woodcock–Johnson III Normative Update (WJ-III NU).

were calculated by the WJ-III NU computer scoring program using either age- or grade-based norms. Specifically, we were interested not only in the standard score (and thus percentile rank) changes that might occur when using age- or grade-based

Table 1. Standard Scores for Perfect Performance as a Function of Age- and Grade-Based Norms for Five Subtests of the Woodcock–Johnson III (WJ-III).

	Normative basis for determining standard scores							
	Age (years)							
WJ-III subtest	18	19	20	21	23	25	27	29
Reading Fluency	117	116	116	116	116	116	116	114
Math Fluency	121	119	118	116	115	115	115	116
Letter-Word Identification	125	123	122	121	119	118	116	116
Word Attack	121	121	120	120	119	119	118	118
Basic Reading Skills	128	126	125	124	123	121	120	120
	Grade							
	13	14	15	16	17			
Reading Fluency	114	111	111	110	109			
Math Fluency	117	113	113	111	110			
Letter-Word Identification	120	117	117	115	113			
Word Attack	117	118	119	121	118			
Basic Reading Skills	120	119	120	119	117			

Table 2. Number of Cases in Each Classification as a Function of Norms Used.

	Classification					
	Borderline		Average		Superior	
	Basis for norms used					
WJ-III subtest	Grade	Age	Grade	Age	Grade	Age
Reading Fluency	40	29	30	41	0	0
Math Fluency	23	14	47	55	0	1
Letter-Word Identification	31	15	39	51	0	4
Word Attack	22	0	47	68	1	2
Basic Reading Skills	31	15	39	50	0	5

Note. 14 ability levels × 5 ages = 70 cases in total. WJ-III = Woodcock–Johnson III.

norms but also whether differences in obtained scores changed across the spectrum of possible raw scores or across age versus grade levels. In addition, we wondered to what extent test ceilings (i.e., the highest score possible for any age or grade) might artificially create discrepancies that could be interpreted as indications of a young adult being both gifted and SLD, especially if grade-based scores were employed when determining academic achievement. Finally, we wondered how many identical

scores would be classified differently (i.e., interpreted differently) when grade norms were used for score interpretation as opposed to age-based norms.

Results show clearly that one consistently obtains lower standard scores when applying grade- as opposed to age-level normative data to interpret the meaning of the same raw scores and that the magnitude of this score difference increases both as proficiency level (i.e., number correct) decreases and as age/grade increases. This has significant implications for test data interpretation, especially when evaluating upper year or graduate-level students. Specifically, across all tests examined in this study, the largest differences in calculated scores were obtained in the older students (age 25-29 years) when compared with Grade 17–level norms. While marginal means were used to illustrate the obtained differences, the individual difference between age and grade scores even at the highest age and proficiency levels was often more than 1 standard deviation for subtest scores such as Reading Fluency, Letter-Word Identification, and Basic Reading Skills. Furthermore, as raw score decreased, the magnitude of the age–grade difference increased significantly. Hence, as Brooks et al. (2011) warned, when low-functioning individuals are excluded from a normative sample, the resulting low end of the new distribution (or the lowest percentiles) are now occupied by persons who would have populated the higher percentiles in the normal distribution of all individuals in the general population. This, in turn, leads to otherwise normally functioning individuals being falsely identified as low functioning or impaired. Given that a large proportion of the U.S. population at the time of norming had completed neither Grade 12 nor a 4-year university program, it seems reasonable to conclude that the age-based normative data are likely the best proxy for use when determining impairment relative to most other people in the general population. Use of such norms will also assist in best determining whether a normative impairment exists that requires academic accommodations and, thus, ensure that test accommodations are provided appropriately.

These findings are important in light of the considerable diversity among the methods currently employed to diagnose SLD in young adults. Two individuals of equal ability assessed using age- and grade-based norms, respectively, could receive different diagnoses depending on norms employed, which may result in differential treatment and access to resources or accommodations. An individual assessed using grade norms will consistently score lower than an equal counterpart assessed using age norms.

Even without employing grade-based norms, subtest and cluster score ceilings may inadvertently interfere with accurate diagnosis of SLD, especially when assessing individuals with higher overall intelligence. Despite being widely criticized, many clinicians still employ discrepancy formulae when diagnosing SLD (Gordon et al., 2002; Weiss et al., 2017). If a clinician employed a discrepancy formula for diagnosis of SLD, especially in a higher functioning adult, results from this analysis show that individuals aged 25 to 29 years with an FSIQ of 143 or greater would meet the discrepancy criteria between IQ and Basic Reading Skills on the WJ-III NU as suggested by McCallum et al. (2013) even when the individual achieved a perfect score

on the latter test cluster (i.e., the highest score on Basic Reading that one can receive in this age range is between 120 and 121). If a clinician employed only a single subtest score, individuals in the same age range with FSIQ scores of 132 or higher would be said to have a 15-point IQ–achievement discrepancy, even though the student had achieved a perfect score on Reading Fluency or Math Fluency. Employing grade-based norms exacerbates the problem that occurs when using such discrepancy criteria, resulting in test score ceilings (i.e., perfect scores) that can fall as low as the average range and, thus, allowing for more individuals to demonstrate an illusory IQ–achievement discrepancy (e.g., a 1 standard deviation discrepancy found for someone with an FSIQ of 124 and grade-based Reading Fluency score of 109) even when a perfect score was obtained.

Results from this evaluation show also that test ceilings on some of the WJ subtests mean that few young adults can achieve above-average scores and that applying grade-based norms to interpret raw scores decreases the number who would be considered to have average or better achievement and increases the number classified as impaired. This finding again supports the contentions made by Brooks et al. (2011) regarding the inappropriateness of using grade-based norms to determine normative impairment.

Limitations

The main limitation of this study was the use of the WJ-III NU norms as opposed to the more recent norms offered by the WJ-IV, published in 2014. The WJ-IV sample is slightly smaller, but it still provides both age and grade norms. While it would have been preferable to also explore the score differences obtained using this newer version of the test, this was not economically feasible; instead of being able to score unlimited test protocols as was true for the WJ-III NU, the newer version allows for online scoring only, with the user having to pay an additional fee for each scored protocol. Given the number of calculations required in the present study, this was simply not possible. We have no reason to believe, however, that the differences identified in the present study would be substantially lower in the newer version of this test.

Conclusions

To determine whether a young adult is impaired relative to most other individuals, results from this analysis show clearly that one must apply age-based normative data when interpreting the meaning of any obtained raw score. This recommendation is in line with that of the ADAAA, which specifies that one must determine whether a person is substantially impaired relative to most other people in the general population when evaluating disability status. Furthermore, disability services offices at postsecondary institutions and high-stakes testing agencies should require that clinicians disclose the normative criteria by which the reported achievement results were determined. In that way, these decision makers can verify that an individual requesting academic accommodations has a normative rather than a relative weakness.

When, then, might one want to employ grade-based norms in an assessment? If the purpose of the assessment is to determine how an individual is performing relative to others at a given level of educational attainment (e.g., ranking how well one plays golf compared with prospective collegiate-level players), then the grade-based scores from the WJ may be appropriate. There may be times when one wishes to know why a student is struggling in a certain program or course, and a ranking within the elite levels may demonstrate that the student's skills may not be sufficient for the task. For individuals attempting to play collegiate-level golf, knowing that their otherwise normal golfing abilities do not measure up to the elite competitors they now face may be important information when making decisions regarding career prospects. Similarly, understanding that a student's otherwise normal achievement scores are nevertheless lower than others in a given graduate-level program may provide useful information about the potential reasons why the student experiences difficulties understanding the material being taught.

If, however, the purpose is to identify a normative impairment relative to most other individuals in the general population, then one needs to determine how the person is performing relative to that broad reference group. If your purpose were to determine, for instance, if an individual is ''golf disabled,'' one would need to compare her skills to all golfers to see if she is unable to do what ''the majority'' of golfers can achieve. The grade-based postsecondary norms offered by the WJ fail to provide this type of normative comparison when evaluating academic achievement in young adults.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Partial funding for this research was provided by the Ministry of Training, Colleges and Universities of Ontario. The opinions as expressed in this paper are those of the authors and do not necessarily reflect those of the funders.

ORCID iD

Allyson G. Harrison https://orcid.org/0000-0002-0426-2011

References

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Americans with Disabilities Act Amendments 2008, 42 U.S.C. §12101 *et seq*. (2008).

Barber, C., & Mueller, C. T. (2011). Social and self-perceptions of adolescents identified as gifted, learning disabled, and twice-exceptional. *Roeper Review*, *33*, 109-120. doi: 10.1080/02783193.2011.554158

Bolt, S. E., & Thurlow, M. L. (2004). Five of the most frequently allowed testing accommodations in state policy. *Remedial and Special Education*, *25*, 141-152.

Brooks, B., Sherman, E., Iverson, G. Slick, & Strauss, E. (2011). Psychometric foundations for the interpretation of neuropsychological test results. In M. R. Schoenberg & J. G. Scott (Eds.). *The little black book of neuropsychology: A syndrome-based approach* (pp. 893-922). New York, NY: Springer.

Cox, M. L., Herner, J. G., Demczyk, M. J., & Nieberding, J. J. (2006). Provision of testing accommodations for students with disabilities on statewide assessments: Statistical links with participation and discipline rates. *Remedial and Special Education*, *27*, 346-354.

Cressman, M. N., & Liljequist, L. (2014). The effect of grade norms in college students: Using the Woodcock–Johnson III Tests of Achievement. *Journal of Learning Disabilities*, *47*, 271-278.

Deloitte, LLP and Affiliated Entities. (2017). *Enabling sustained student success: Support for students at risk in Ontario's colleges*. Retrieved from http://www.collegesontario.org/policy-positions/position-papers/Colleges%20Ontario%20_At%20Risk%20Student%20Report_2017_vfinal.pdf

Fuchs, L. S., Fuchs, D. (2001). Principles for the prevention and intervention of mathematics difficulties. *Learning Disabilities Research & Practice*, *16*, 85-95. doi:10.1111/0938-8982.00010

Gilman, B. J., Lovecky, D. V., Kearney, K., Peters, D. B., Postma, M., Wasserman, J. D., . . . Rimm, S. B. (2013). Critical issues in the identification of gifted students with co-existing disabilities the twice-exceptional. *SAGE Open*, *3*(3). doi:10.1177/2158244013505855

Giovingo, L. K., Proctor, B. E., & Prevatt, F. (2005). Use of grade-based norms versus age-based norms in psychoeducational assessment for a college population. *Journal of Learning Disabilities*, *38*, 79-85.

Gordon, M., Lewandowski, L., Murphy, K., & Dempsey, K. (2002). ADA-based accommodations in higher education: A survey of clinicians about documentation requirements and diagnostic standards. *Journal of Learning Disabilities*, *35*, 357-363.

Gregg, N., Coleman, C., Lindstrom, J., & Lee, C. (2007). Who are most, average, or high-functioning adults? *Learning Disabilities Research & Practice*, *22*, 264-274. doi: 10.1111/j.1540-5826.2007.00255.x

Gyenes, J., & Siegel, L. S. (2014). A Canada-wide examination of the criteria employed for learning disability documentation in English speaking postsecondary institutions. *Canadian Journal of School Psychology*, *29*, 279-295.

Harrison, A. G. (2017). Clinical, ethical, and forensic implications of a flexible threshold for LD and ADHD in postsecondary settings. *Psychological Injury and Law*, *10*, 138-150.

Harrison, A. G., & Holmes, A. (2012). Easier said than done: Operationalizing the diagnosis of learning disability for use at the postsecondary level in Canada. *Canadian Journal of School Psychology*, *27*, 12-34.

Harrison, A. G., Lovett, B., & Gordon, M. (2013). Documenting disabilities in postsecondary settings: Diagnosticians' understanding of legal regulations and diagnostic standards. *Canadian Journal of School Psychology*, *28*, 303-322.

Harrison, A. G., & Wolforth, J. (2012). Findings from a pan-Canadian survey of disability services providers in postsecondary education. *International Journal of Disability, Community and Rehabilitation*, *11*(1). Retrieved from http://www.ijdcr.ca/VOL11_01/articles/harrison.shtml

Iverson G., L., & Brooks, B. L. (2011). Improving accuracy for identifying cognitive impairment. In M. R. Schoenberg & J. G. Scott (Eds.). *The little black book of neuropsychology: A syndrome-based approach* (pp. 923-950). New York, NY: Springer.

Julian, E. R. (2005). Validity of the medical college admission test for predicting medical school performance. *Academic Medicine*, *80*, 910-917.

Keiser, S. (1998). Test accommodations: An administrator's view. In M. Gordon & S. Keiser (Eds.), *Accommodations in higher education under the Americans with Disabilities Act*. New York, NY: Guilford Press.

Kettler, R. (2012). Testing accommodations: Theory and research to inform practice. *International Journal of Disability, Development and Education*, *59*, 53-66.

Kimball, E. W., Wells, R. S., Ostiguy, B. J., Manly, C. A., & Lauterbach, A. A. (2016). Students without disabilities in higher education: A review of the literature and an agenda for future research. In m. B. Paulsen (Ed.), *Higher education: Handbook of theory and research* (pp. 91-156). Cham, Switzerland: Springer International.

Lai, S. A., & Berkeley, S. (2012). High-stakes test accommodations: Research and practice. *Learning Disability Quarterly*, *35*, 158-169.

Lauth, L., Sweeney, A., & Reese, L. (2017). *Accommodated test-taker trends and performance for the June 2012 through February 2017 LSAT administrations*. Retrieved from https://www.lsac.org/data-research/research/accommodated-test-taker-trends-and-performance-june-2012-through-february

Lerner, C. (2004). ''Accommodations'' for the learning disabled: A level playing field or affirmative action for elites? *Vanderbilt Law Review*, *57*, 1041-1122.

Lewandowski, L. J., Codding, R. S., Kleinmann, A. E., & Tucker, K. L. (2003). Assessment of reading rate in postsecondary students. *Journal of Psychoeducational Assessment*, *21*, 134-144.

Lewandowski, L. J., Cohen, J. A., & Lovett, B. J. (2013). Effects of extended time allotments on reading comprehension performance of college student with and without learning disabilities. *Journal of Psychoeducational Assessment*, *31*, 326-336.

Lindstrom, W., & Lindstrom, J. H. (2017). College admissions tests and LD and ADHD documentation guidelines. *Journal of Disability Policy Studies*, *28*, 32-42. doi:10.1177/1044207317696261

Lovett, B. J. (2010). Extended time testing accommodations for students with disabilities: Answers to five fundamental questions. *Review of Educational Research*, *80*, 611-638.

Lovett, B. J., Gordon, M., & Lewandowski, L. J. (2009). Measuring impairment in a legal context: Practical considerations in the evaluation of psychiatric and learning disabilities. In S. Goldstein & J. Naglieri (Eds.), *Assessing impairment: From theory to practice* (pp. 93-103). New York, NY: Springer.

Madaus, J. W., Banerjee, M., & Hamblet, E. C. (2010). Learning disability documentation decision making at the postsecondary level. *Career Development for Exceptional Individuals*, *33*, 68-79.

Maddocks, D. L. (2018). The identification of students who are gifted and have a learning disability: A comparison of different diagnostic criteria. *Gifted Child Quarterly*, *55*, 3-17.

McCallum, R. S., Bell, S. M., Coles, J., Miller, K., Hopkins, M., & Hilton-Prillhart, A. (2013). A model for screening twice-exceptional students (gifted with learning disabilities) within a Response to Intervention paradigm. *Gifted Child Quarterly*, *57*, 209-222. doi: 10.1177/0016986213500070

McCoach, D. B., Kehle T., Bray, M. A., & Siegle D. (2001). Best practices in the identification of gifted students with learning disabilities. *Psychology in the Schools*, *38*, 403-411.

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). Handbook of normative data for neuropsychological assessment (2nd ed.). Oxford, England: Oxford University Press.

Ofiesh, N. S., Hughes, C., & Scott, S. S. (2004). Extended test time and postsecondary students with learning disabilities: A model for decision making. *Learning Disabilities Research and Practice*, *19*, 57-70. doi:10.1111/j.1540-5826.2004.00090.x

Psychological Corporation. (1997). *WAIS-III and WMS-III technical manual*. San Antonio, TX: Harcourt Brace.

Ranseen, J. D., & Parks, G. S. (2005). Test accommodations for postsecondary students: The quandary resulting from the ADA's disability definition. *Psychology, Public Policy, & Law*, *11*, 83-108.

Ready, R. E., Chaudhry, M. F., Schatz, K. C., & Strazzullo, S. (2012). ''Passageless'' administration of the Nelson-Denny Reading Comprehension Test: Associations with IQ and reading skills. *Journal of Learning Disabilities*, *46*, 377-384.

Roberts, B. (2012). Beyond psychometric evaluation of the student—task determinants of accommodation: Why students with learning disabilities may not need to be accommodated. *Canadian Journal of School Psychology*, *27*, 72-80. doi: 10.1177/0829573512437171

Schrank, F. A., Mather, N., & McGrew, K. S. (2014). *Woodcock-Johnson IV Tests of Achievement*. Rolling Meadows, IL: Riverside.

Stretch, L. S., & Osborne, J. W. (2005). Extended time test accommodation: Directions for future research and practice. *Practical Assessment, Research, and Evaluation*, *10*(8), 1-8.

Thomas, S. B. (2000). College students and disability law. *Journal of Special Education*, *33*, 248-257.

Thurlow, M. L., Thompson, S. J., & Lazarus, S. S. (2006). Considerations for the administration of tests to special needs students: Accommodations, modifications, and more. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 653-673). Mahwah, NJ: Lawrence Erlbaum.

U.S. Census Bureau. (2016). *Educational attainment in the United States: 2016*. Retrieved from https://www.census.gov/data/tables/2016/demo/education-attainment/cps-detailed-tables.html

Weiss, R., Erikson, C., & Till, C. (2017). When average is not good enough: Students with learning disabilities at selective, private colleges. *Journal of Learning Disabilities*, *50*, 684-700. doi:10.1177/0022219416646706

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001a). *Woodcock-Johnson–III Tests of Achievement*. Itasca, IL: Riverside.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001b). *Woodcock-Johnson–III Tests of Cognitive Abilities*. Itasca, IL: Riverside.

Woodcock, R. W., McGrew, K. S., Schrank, F. A., & Mather, N. (2007). *Woodcock-Johnson–III Normative Update*. Rolling Meadows, IL: Riverside.

Yellin, D. (2016). *Growing number of students seeking accommodations for SAT*. Retrieved from http://www.northjersey.com/story/news/education/2017/01/26/record-numbers-students-seeking-accommodations/96162464/