Psychological Injury and Law (2019) 12:17–28
https://doi.org/10.1007/s12207-019-09340-2

A Critical Analysis of the Nelson Denny Reading Test as a Method of Identifying Reading Impairment in Adults

Allyson G. Harrison[1] · Kathleen A. Harrison[2]

Received: 4 January 2019 / Accepted: 10 January 2019 / Published online: 2 February 2019

Abstract

Disability-related test accommodations are requested frequently, especially at the postsecondary level and on licensing examinations. Access to such accommodations typically relies on proof of impairment in some area of academic functioning. The Nelson Denny Reading Test (NDRT; Brown, Fishco, & Hanna, 1993a, 1993b) is often employed by clinicians in order to demonstrate the need for extra time accommodation. The NDRT employs grade-based norms, meaning that postsecondary and graduate-level students who take the test are compared not with all of their same-aged peers but rather to a rarefied group of individuals who have achieved equally high levels of education. This leads to a skewed distribution of scores that, in turn, makes otherwise normally functioning individuals appear impaired. Employing the actual normative data from the NDRT, this study investigated the effect that use of such grade-based norms has on ratings of normative and relative impairment. With the same raw score, substantially more individuals would be classified as impaired on a measure of timed reading comprehension when higher grade level norms are applied as compared with norms that represent a broader sample of individuals. These findings demonstrate clearly that grade-based norms should not be employed when using the NDRT to determine disability-related normative impairment.

Keywords Assessment · Norms/normative · Disability · Reading · Adult

Academic testing accommodations allow individuals with disabilities to participate equally when construct-irrelevant variables would otherwise interfere with accurate evaluation of knowledge or skills (Fuchs & Fuchs, 2001; Lai & Berkeley, 2012; Roberts, 2012; Thurlow, Thompson, & Lazarus, 2006). When used properly, testing accommodations lead to more valid test scores for examinees with disabilities; if given inappropriately however they may give examinees an unfair advantage allowing them to access significantly more of the test material than their non-disabled peers (Lerner, 2004; Lewandowski, Cohen, and Lovett, 2013; Lovett, 2010; Phillips, 1994). As such, accurate measurement of normative impairment is essential when making disability determinations

✉ Allyson G. Harrison
harrisna@queensu.ca

[1] Regional Assessment & Resource Centre, Queens University, Kingston, ON, Canada

[2] Centre for Neuroscience Studies at Queen's University, Kingston, ON, Canada

The most commonly requested accommodation is extra time (Bolt & Thurlow, 2004; Harrison & Wolforth, 2012; Julian et al., 2004; Keiser, 1998; Kettler, 2012; Lewandowski et al., 2013; 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 (that is, completed in the past 3–5 years; Gyenes & Siegel, 2014; Lindstrom & Lindstrom, 2017).

High-stakes testing agencies such as Educational Testing Services (2017) indicate that the need for extra test time must be verified using a reliable, valid, and standardized test for use with an adolescent or adult. The evaluator must use standard scores if possible, and document a substantial limitation relative to most other individuals in the general population. Unfortunately, very few standardized measures of reading speed or timed reading comprehension ability exist with appropriate adult normative data.

One of the tests most commonly employed to demonstrate impaired speed of reading or comprehension in young adults is the Nelson Denny Reading Test (NDRT; Brown et al., 1993a). It is a timed test of reading comprehension and

vocabulary, and offers a 1-min reading speed measure as well. Two parallel forms allow for test-retest comparison. The scoring manual allows test administrators to calculate scale scores in relation to obtained raw scores for each of these three subtests. Test users may compare this scale score with normative data for individuals in various school grades ranging from grade 9 through the end of grade 16 (i.e., end of a 4-year university degree).

Although the test manual warns users that obtained scores are not to be used for purposes of diagnosis, in absence of any other standardized measures, many clinicians employ the NDRT for just this purpose (Lewandowski et al., 2013; Ready, Chaudhry, Schatz, & Strazzullo, 2012). Further, use of the NDRT as proof of the need for extra time accommodations is recommended by many high-stakes testing agencies (Ready et al., 2012) and clinicians (e.g., Kettler, 2012; Ofiesh et al., 2004).

Unfortunately, it appears that the NDRT has many identified limitations to its usefulness as part of a diagnostic assessment, especially when assessing students in undergraduate, graduate, or postgraduate programs. When evaluated by the Mental Measurements Yearbook in 1998, two independent reviewers both identified that the NDRT had many psychometric limitations and stressed that there was no empirical support for its use as a diagnostic measure (Smith, 1998; Murray-Ward, 1998). They noted that the standardization sample was skewed towards more highly educated White individuals from middle socioeconomic status levels and not representative of the socioeconomic or racial makeup of the USA's population at the time. Content validity of the comprehension section was said to be questionable, the technical manual lacked data regarding predictive or concurrent validity, and the test manual warned users not to overinterpret test scores. The reviews noted that the grade equivalent scores provided in the manual were misleading and unreliable. This is especially true at higher grades (see commentary below). Both reviewers concluded that the NDRT was appropriate only as a crude screening measure and lacked validity as a method of diagnosing reading problems. This reechoed the conclusions of Van Meter and Herrmann (1986) who reviewed the previous version of the NDRT. Despite such criticisms, however, the NDRT is still one of the most widely used tests to document impairments in speed of reading comprehension in order to support need for academic accommodations.

Determining Disability

According to the Americans with Disabilities Act, as amended (ADAAA, 2008), an impairment rises to the level of a disability if it substantially limits the ability of an individual to perform a major life activity as compared to most people in the general population. In general, courts have interpreted this to mean that the individual must be performing substantially below average relative to most other people their age (e.g., *Bibber v. National Board of Osteopathic Medical Examiners, Inc.*, 2016), a conclusion supported by diagnostic criteria for many cognitive disorders listed in the *Diagnostic and Statistical Manual of Mental Disorders*, Fifth Edition (DSM-5; American Psychiatric Association, 2013). As such, extra time is a disability-related accommodation that should be provided only to postsecondary students who have a documented functional impairment that keeps them from being able to complete a test as quickly as most other people in the general population. The question, however, is whether the NDRT can identify functional impairments that require extra time accommodation, especially in relation to the ADAAA standard.

In order to determine whether an individual is "substantially impaired" relative to most other people in the general population, one must employ tests that compare the test taker to this same normative population. Unlike almost all other achievement tests that obtain a broad normative sample from most individuals in the general population, the NDRT offers only grade-based normative data. This means that the test taker is being compared not to all other individuals at a given age but to only those few individuals who attained a certain level of higher education. The normative data for the NDRT were collected in 1991 and 1992 using the 1980 US Census as a guide for sampling characteristics (Brown et al., 1993a). According to the US Census Bureau, only 80.2% of persons age 25 or older had obtained even a *high school* diploma in 1993, and only 21.9% of this group had completed a bachelor's degree or higher (U.S. Census Bureau, 2017). Hence, concerns arise regarding the validity of using such grade-specific norms to determine functional impairment in undergraduate or graduate-level students in relation to most other individuals in the general population, as most people at that time did not participate in or complete higher education degrees.

Brooks, Sherman, Iverson, Slick, and Strauss (2011) reviewed the significant problems that can occur when clinicians employ tests with 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 tennis skills by comparing her not to all female tennis players her age but to only those few who played collegiate-level tennis. Relative to such an elite comparison group, an otherwise average tennis player would be classified

as impaired. Brooks et al. (2011) showed that when low-functioning individuals are excluded from a normative sample, the resulting low end of the new distribution (or lowest percentiles) are now occupied by a person who would have populated the higher percentiles in the normal distribution of all individuals in the general population. They warn that this, in turn, can lead to otherwise normally functioning individuals being falsely identified as low functioning or impaired.

Prior studies (Cressman & Liljequist, 2014; Giovingo, Proctor, & Prevatt, 2005) have shown that this type of misclassification can occur in tests other than the NDRT when using grade-based norms to determine whether a student has a Learning Disability (LD). Indeed, both studies cited above showed clearly that a much higher proportion of postsecondary students would be classified as LD when using grade- as opposed to age-based norms. In the Giovingo et al. study, proportions of symptomatic students whose achievement scores fell below the 16th percentile changed from 40.6% when employing age-based norms to 66.5% when using grade-based norms. Cressman and Liljequist's findings were even more striking. Using an obtained score below the 16th percentile as a proxy for academic impairment, they found that choice of normative comparison group had an extreme impact; whereas between 5.9 and 13.4% of postsecondary students in their sample obtained an achievement score below this level compared with their same-aged peers, between 33.2 and 55.9% fell below this level in comparison to their same-grade peers.

Given that the NDRT normative sample contained mainly middle- to high-income individuals, especially in the 4-year university samples, and that very few adults at that time completed 4 years of university, one must question the extent to which scores obtained when using grade-based norms from the NDRT might inaccurately identify normally functioning students as being impaired.

This study therefore examined the NDRT normative data and technical manual to answer the following questions: First, to what extent do identical scaled scores from Reading Rate and Reading Comprehension change in relative performance value (i.e., percentile ranking) depending on grade level of norms applied, and to what extent does this alter interpretation of the score value (i.e., impaired or not)? (Typically, students in upper grades are given only the comprehension and reading rate tests and not the vocabulary subtest. Hence, we examined only these two subtests.) We hypothesized that individuals whose scores would otherwise be considered normally achieving relative to most other individuals in the general population would now be classified as impaired when their raw scores were interpreted using third- or fourth-year university grade-based norms from the NDRT, and that fewer students in upper grades could achieve scores classified as being average or above average. Second, and relatedly, we reviewed mean scores at each grade level from the technical manual to determine how these average scores were ranked using both same-grade level norms and the pooled normative sample. Also of interest was the number of students at each grade level who were able to finish all or 75% of the comprehension subtest in the normal time provided. Finally, we investigated which reference group from the NDRT appears to best represent a proxy for most people in the general population. This involves evaluating size of each grade-based sample and the proportion of scores in each grade that could be classified as being above or below average.

Method

Materials

In this study, we employed the normative data provided in the scoring manual for the NDRT (Brown et al., 1993a), Form G, Reading Comprehension and Reading Rate. The use of the actual normative data provides a particularly useful frame of reference from which to evaluate all possible scores across the NDRT grade spectrum. In most studies, pre-selected samples of individuals are drawn from clinical samples, which likely produce a restriction of range. In this study, instead, we used the actual normative data, thus avoiding the potential confound of sample pre-selection effects on the dependent variables of interest.

The scoring manual provides all possible raw scores for the Reading Comprehension and one-minute Reading Rate subtests of the NDRT. The Reading Comprehension subtest consists of seven passages with a total of 38 comprehension questions based on these passages. Examinees are given 20 min to read the passages silently and answer factual or inferential multiple-choice questions about the material. Reading Rate is determined when examinees are stopped after 1 min and asked to self-identify what line of the passage they are reading at that moment. There are 41 lines of text allowing for 41 possible scores. The NDRT Technical Manual (Brown et al., 1993b) explains that the length of the first passage was deliberately chosen such that almost no student, regardless of grade level, could complete it in 1 min.

Other information of interest was provided only in the NDRT Technical Manual (Brown et al., 1993b). This included information regarding size of the normative groups in each grade category, and number of students in each age group who were able to complete 75 or 100% of the Comprehension questions within the allotted standard time limit. The technical manual is not provided when the NDRT is purchased, and is currently out of print. As such, it seems likely that many test users have not reviewed the information contained in this separate document.

Procedure

The range of all possible scores for both Reading Comprehension and Reading Rate were calculated for form G of the NDRT. Possible raw scores for the Comprehension test ranged from 0 to 38, which translates into corrected raw scores from 0 to 76 (the manual instructs the test user to double the obtained raw score before interpretation). For Reading Rate, the manual offers 41 possible raw scores ranging from 7 to 601 (i.e., the story has 41 lines, such that finishing the first line produces a score of 7 words read and finishing all 41 lines produces a score of 601). All possible raw scores for both Reading Rate and Comprehension were recorded and then converted into scaled scores as outlined in the manual for scoring and interpretation (Brown et al., 1993a). Each possible raw score corresponds with only one scaled score, and the relative meaning of each scaled score changes depending on grade level chosen for interpretation. As such, the associated percentile scores for each scaled score was then determined relative to the normative data provided for seven different grade levels: students in Grade 12 (senior year high school), grades 13 and 14 from 2-year colleges, and grades 13 through 16 from 4-year universities. All percentile scores were calculated relative to the end-of-year norms provided in the test manual. This process yielded seven different percentile scores for each calculated scaled score (one for each grade cohort). Given, however, that percentile scores are not on an interval scale, we converted each obtained percentile score to a normal curve equivalent (NCE) score using the conversion table provided in the NDRT manual for scoring and interpretation (p. 41). NCE scores are standardized scores with a mean of 50 and a standard deviation of 21.06, and these scores represent equal units. As such, they may be employed to compute averages and make statistical comparisons regarding score changes obtained when using different grade-based norms.

As an alternative to using grade-based norms for interpretation, the test manual notes that the scaled scores provided in the normative manual are "based on the pooled standardization sample from Grades 10, 11, and 12 and both two-year college classes and both lower division classes in the four-year institutions" (p. 14). The unweighted pooled distribution from these seven different groups is combined into a system of normalized scaled scores for which the mean was set at 200 with a standard deviation of 25. The size of this pooled normative sample was 8600. By contrast, the sizes of the samples in the upper-year university normative groups were relatively small (see Table 1).

To investigate the magnitude of score change obtained when identical scale scores are interpreted using different grade-based norms, we then tabulated the various percentile (and then NCE) scores returned when identical data were interpreted using norms for each of Grades 12 through 16, and compared these to either the pooled scale scores from the test manual or to individuals at the end of high school (assuming that one or the other of these would more closely resemble "most people in the general population"). To allow for a comparison between ranking using grade norms and the pooled norms, scale scores were converted to NCE scores.

Table 1 NDRT normative sample sizes by grade

	N
Pooled standardization sample size	8600
Grade 10	1714
Grade 11	1483
Grade 12	1271
Grade 13 2-year college	2023
Grade 14 2-year college	1043
Grade 13 4-year university	1043
Grade 14 4-year university	488
Grade 15 4-year university	584
Grade 16 4-year university	558

Results

Difference scores across the spectrum of Reading Comprehension-derived scores were calculated by subtracting the corresponding NCE values between scores generated when using the Grade 12 norms and scores obtained when using grade-based normative data from Grade 13 and higher. These difference scores were then converted to z scores. As shown in Fig. 1, obtained NCE scores decreased across all but the lowest scores or the very highest two or three scores. Further, a trend was evident in that calculated NCE scores (and associated percentile scores) decreased as comparison group increased in grade level. In fact, the same raw score is systematically ranked lower as grade level increases, such that if Grade 16 norms are employed there was just over a one-SD decline in calculated scores across the majority of the distribution when compared with rankings obtained using the Grade 12 norms. The only exceptions were at the lowest end (raw scores below 8 which translated consistently into a score at the first percentile), raw scores between 9 and 12 (where the NCE difference was slightly less than 1 SD), and perfect or near perfect scores (raw scores of 38 or 37 which also yielded consistent scores at the 99th percentile). Even using Grade 15 norms resulted in a NCE score decrease of half to three-quarters of a standard deviation across most possible raw scores.

A similar pattern was obtained if the difference scores were derived using pooled standardization norms as a proxy for "most people in the general population" and comparing these to the grade-based normative data from Grade 13 and higher. As shown in Fig. 2, the pattern of

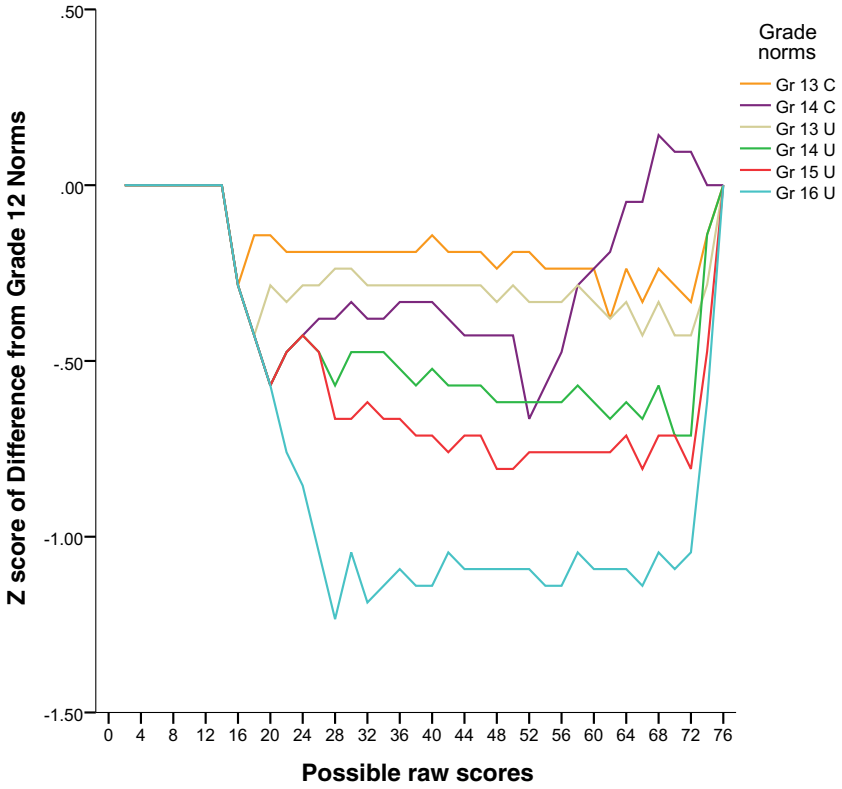

Fig. 1 Z score of the difference in Reading Comprehension scores between the Grade 12 norms and grade levels as a function of all possible raw NDRT scores

score decline remained (i.e., similar raw scores consistently result in lower NCE scores as grade level increased), but the magnitude of the difference was even greater. Indeed, both Grade 14 and 15 norms now consistently yielded score differences of close to 1 SD across most possible scores, and the grade 16 norms changed the

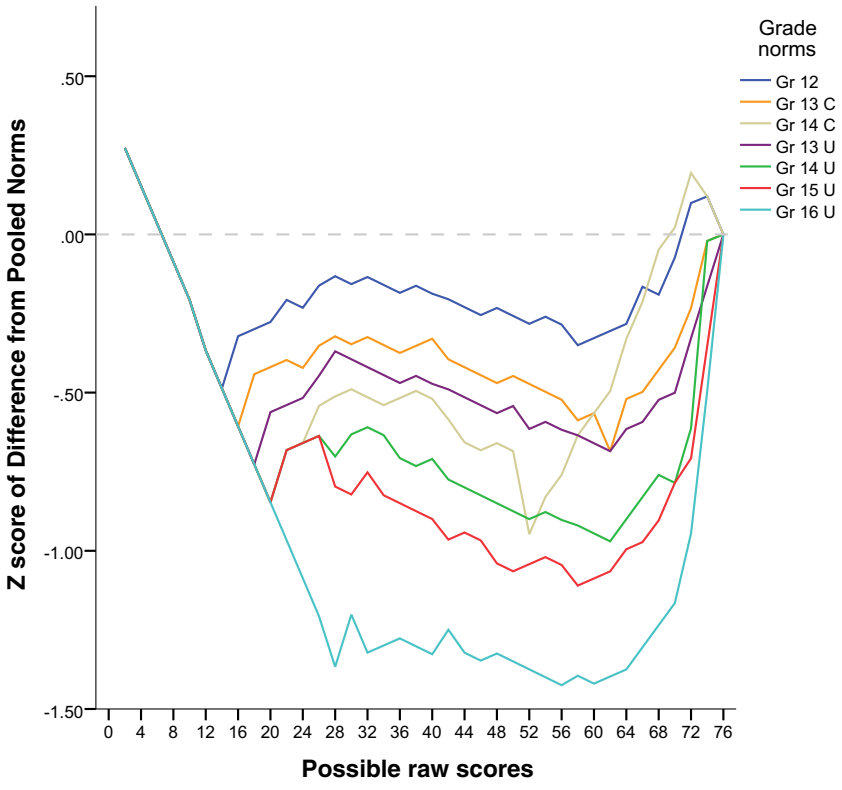

Fig. 2 Z score of the difference in Reading Comprehension scores between pooled norms and each grade level as a function of all possible raw NDRT scores

interpretation by almost 1.5 SD across most of the possible raw score distribution.

Also of interest was the change in classification of scores (average, 16th to 84th percentile; above average, above the 84th percentile; or impaired, below the 16th percentile) depending on which set of norms was employed to interpret identical raw scores. For timed Reading Comprehension, it was clear that as grade level increased, the number of scores interpreted as impaired increased and the number said to be average or above average decreased (see Table 2). In fact, using norms from Grade 16 only two scores (5.3%) would be classified as above average, 13.6% classified as average, and 63.2% of all possible scores are interpreted as impaired. Only when one employs the pooled norms to interpret the meaning of similar raw scores does one obtain a distribution that even remotely resembles a normal distribution; even here, however, more scores are interpreted as falling below as compared to above average. This means that the normative sample on which these scores are based is skewed, having relatively higher functioning and fewer lower functioning individuals in the sample.

No consistent pattern was found when calculating Reading Rate score differences between Grade 12 and other grades. As shown by Fig. 3, while identical raw scores were interpreted as falling at a lower NCE level as grade level increased, this occurred primarily at the lower raw score level.

The trend in classification change was not observed when examining the distribution of possible Reading Rate scores. Indeed, as shown in Table 3, across all grade levels, more raw scores are considered to fall above average than in any other classification. Scores interpreted as being average outnumber those considered impaired only when the pooled norms or Grade 12 norms are used to interpret obtained raw scores. Using any other grade level normative data, more scores are interpreted as being impaired relative to average. At the Grade 16 level, 31.7% of all possible scores are classified as impaired, whereas only 22% of these same scores would be so classified by the pooled norms.

Given that the technical manual does not provide information regarding the score distributions at each grade or how percentile rankings were calculated by grade, it was of interest to investigate how the average student performed in each normative grade cohort. One might assume, for instance, that an average score at a given grade level would translate into a score at the 50th percentile, and that other scores would be distributed evenly around this mean. We therefore used the mean raw scores and standard deviations (SD) data provided in the technical manual (Brown et al., 1993b), calculating the associated scaled scores for both Reading Comprehension and Reading Rate in each of the grade levels sampled, and the percentile rank that would be associated with each of these mean scores if the consonant grade level norms were applied. These percentiles show how an average student would be ranked relative to his or her same-grade peers. As may be seen in Table 4, if the scaled scores are used to interpret performance, one finds the expected improvement in average level of functioning relative to most other people in the general population as grade level increases. In other words, average performance improves as grade goes up relative to a broad normative sample. By contrast, the average raw Reading Comprehension score for each grade translates into percentile scores that become progressively smaller at increasing grades, such that the average score obtained by those in the Grade 16 cohort is ranked as being better than only 31% of students at that grade level. Even in Grade 12, the mean score translates into a percentile rank that is slightly lower than average (46th percentile). Further, the standard deviation associated with the mean scores indicates that the score distribution was not normal. Indeed, while a score that is 2 SD below average consistently translates into a score falling in the 1st to 4th percentile (recall that the raw scores range from 0 to 38 and are then doubled before calculating a scaled score), the upper range (i.e., 2 SD above the mean) of all but the Grade 12 scores exceeds the maximum raw score possible.

By contrast, the mean scores obtained from the technical manual for Reading Rate consistently translate into both scaled scores and percentile scores that are just above average, ranging from the 52nd to the 58th percentile. In other words, even using scale scores the average Reading Rate does not change across grade level. Additionally, the standard deviation of Reading Rate scores suggests a non-normal distribution as the upper limit of the distribution is not reached. As shown in Table 5, while a score of 2 SD below the mean results in percentile ranks at the 1st percentile, a score of 2 SD above the mean only equates to a score better than 95 or 96% of the sample, a rank that is less than 2 SD within a normal (Gaussian) distribution of scores. Given that the maximum possible raw score is 601, scores in each grade that are 2

Table 2 Number and percentage of all possible calculated scores falling within, below, or above average for Reading Comprehension on the NDRT

	Below average		Average		Above average	
Norms used	*N*	%	*N*	%	*N*	%
Pooled	13	34.2	16	42.1	9	23.7
Gr. 12	14	36.8	18	47.4	6	15.8
Gr. 13 Two-year college	16	42.1	17	44.7	5	13.8
Gr. 14 Two-year college	17	44.7	15	39.5	6	15.8
Gr. 13 Four-year university	17	44.7	17	44.7	4	10.5
Gr. 14 Four-year university	19	50.0	16	42.1	3	7.9
Gr. 15 Four-year university	21	55.3	14	36.8	3	7.9
Gr. 16 Four-year university	24	63.2	12	13.6	2	5.3

Gr. grade

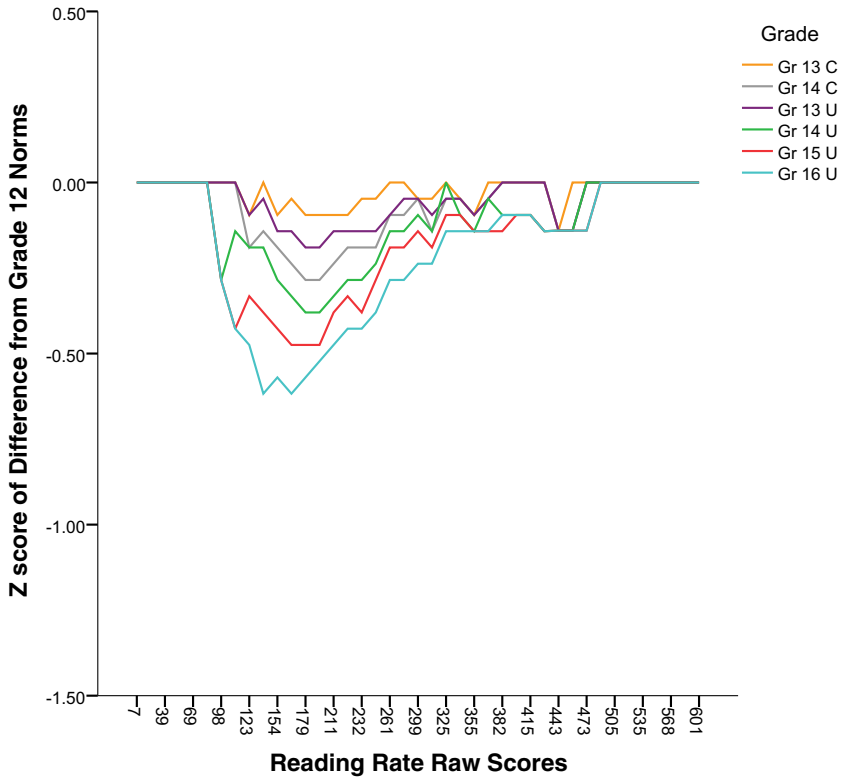

Fig. 3 Z score of the difference in Reading Rate scores between the Grade 12 norms and grade levels as a function of all possible raw NDRT scores

SD above the mean still fall at least 2 SD below the highest possible raw score, again indicating that the distribution of scores is skewed.

Finally, we were interested in determining the percentage of students for whom use of different grade level norms would change the classification of identical obtained raw scores on the timed Reading Comprehension subtest. In other words, how many identical raw scores would change classification level depending on norms employed, assuming either that the Grade 12 or the pooled norms are most representative of the general population. As shown in Table 6, relatively few scores change classification if one uses grade 13 2-year college as opposed to Grade 12 norms. Similarly, relatively few score classifications

Table 3 Number and percentage of all possible scores falling within, below, or above average for Reading Rate on the NDRT

	Below average		Average		Above average	
Norms used	*N*	%	*N*	%	*N*	%
Pooled	9	22.0	12	29.3	20	48.8
Gr. 12	10	24.4	11	26.8	20	48.8
Gr. 13 Two-year college	11	26.8	10	24.4	20	48.8
Gr. 14 Two-year college	12	29.3	10	24.4	19	46.3
Gr. 13 Four-year university	11	26.8	11	26.8	19	46.3
Gr. 14 Four-year university	12	29.3	10	24.4	19	46.3
Gr. 15 Four-year university	12	29.3	10	24.4	19	46.3
Gr. 16 Four-year university	13	31.7	9	22.0	19	46.3

Gr. grade

Table 4 Reading Comprehension mean raw values and SD by grade converted into scaled scores and percentiles for average performance, 2 SD below, and 2 SD above average as a function of grade

Grade	12	13 C	14 C	13 U	14 U	15 U	16 U
Average values							
Mean raw score	45.22	46.78	50.13	48.50	52.44	55.70	59.07
SD	15.74	15.61	15.42	15.27	14.84	14.46	12.29
SS	204	204	210	207	216	219	223
Percentile	46	39	38	39	40	38	31
Values 2 SD below average							
Mean raw scores	13.74	15.56	19.29	17.96	22.76	28.76	34.49
SS	154	157	160	160	166	176	185
Percentile	1	1	1	1	2	4	3
Values 2 SD above average (maximum raw score is 76)							
Raw scores	76.70	78.00	80.39	79.04	82.12	84.62	83.65
SS	251	251	251	251	251	251	251
Percentile	99	99	99	99	99	99	99

All mean scores copied from the NDRT Technical Manual (Brown et al., 1993b)

C 2-year college, *U* 4-year university, *SS* scaled score

Table 5 Reading Rate mean raw values and SD by grade converted into scaled scores and percentiles for average performance, 2 SD below, and 2 SD above average as a function of grade cohort

Grade	12	13 C	14 C	13 U	14 U	15 U	16 U
Average values							
Mean raw score	226.21	230.21	238.21	234.21	242.50	246.22	250.26
SD	87.38	87.37	86.80	87.18	85.77	84.88	83.32
SS	201	204	204	204	208	208	208
Percentile	52	57	52	54	58	56	52
Values 2 SD below average							
Mean raw score	51.54	55.47	64.61	59.85	70.96	76.46	83.62
SS	151	151	155	151	155	159	159
Percentile	1	1	1	1	1	1	1
Values 2 SD above average (maximum raw score is 601)							
Raw scores	400.97	404.95	411.81	408.57	414.04	415.98	416.90
SS	254	254	259	259	259	259	259
Percentile	95	95	96	96	95	95	95

All mean scores copied from the NDRT Technical Manual (Brown et al., 1993b)

C 2-year college, *U* 4-year university, *SS* scaled score

change if pooled norms are employed to interpret the raw scores instead of Grade 12 norms. By contrast, 26% of scores that would be interpreted as average based on Grade 12 norms would now be classified as impaired if Grade 16 norms were applied and 29% if the pooled normative sample was used as the best estimate of general population performance. Fewer, but a still substantial number of possible scores would also be interpreted differently when upper-year (grade 14 or 15 university) normative data are employed as compared with either Grade 12 or the pooled norms. Such score changes were not as evident in Reading Rate, where use of either pooled or Grade 12 norms did not change score classification significantly. Indeed, fewer than 10% of the scores experienced a change in classification if Grade 16 norms were employed compared with Grade 12 or pooled norms, and fewer than 5% of other grade scores result in a classification change relative to these latter norms.

Discussion

The goal of the present study was to evaluate critically the normative data for the NDRT, specifically investigating how the use of grade norms instead of age norms obtained from the general population might change the identification of impairment in adults who have attained various levels of postsecondary education. Examining the normative data collected from students in Grade 12 and higher, it is clear that substantially more students were recruited for the normative sample from the first 2 years of college compared with the upper years of university. Further, the pooled normative data provided in the NDRT manual, representing the combined and normalized data from all students in Grades 10 through 12 and the first 2 years of both college and university, provides a much larger and more broadly representative comparison group against which to determine overall level of functioning. Given that the vast majority of the population at the time of norming

Table 6 Percentage change in classification for Reading Comprehension scores as a function of the norms employed and comparison group chosen (grade 12 or pooled norms)

	Grade 12 norms Average to below average % change	Pooled norms Average to below average % change
Gr. 12		3
Gr. 13 Two-year college	5	8
Gr. 14 Two-year college	8	11
Gr. 13 Four-year university	8	11
Gr. 14 Four-year university	13	16
Gr. 15 Four-year university	18	21
Gr. 16 Four-year university	26	29

Gr. grade

had not completed a 4-year university program, it seems reasonable to conclude that the pooled normative data (mean of 200, SD of 25) or the grade 12 norms are likely the best proxy for use when determining impairment relative to most other people in the general population.

As hypothesized, we demonstrated that individuals whose performance would otherwise be considered normally achieving relative to most other individuals in the general population are described as impaired when the same raw scores are compared to individuals about to graduate from a 4-year university. In fact, just over 63% of all possible raw scores on the NDRT Reading Comprehension subtest would result in a classification of impaired (i.e., below the 16th percentile) when the end-of-4th-year norms are employed, and just over 55% if compared with end-of-3rd-year norms. At the other extreme, only the top two raw scores on the Reading Comprehension subset (a perfect raw score of 38 or a score of 37) are interpreted as falling above average if Grade 16 norms are employed, and only three raw scores qualify as above average if Grade 14 or 15 norms are applied. In fact, even the average Reading Comprehension score achieved by those in Grade 16 is considered to fall at the 31st percentile relative to other students at that grade level, even though the mean score is almost one standard deviation above the average for the broad, pooled norms. Clearly, the standards for measuring relative performance change as grade level increases, such that anything less than perfect or almost perfect is merely average when compared with those at the end of grade 16.

Using all possible raw scores for the NDRT Reading Comprehension measure, we found that as postsecondary grade level increased, NCE (and thus percentile) scores decreased dramatically relative to the same raw score for all but the most extreme scores. Further, assuming that either the Grade 12 normative data or the pooled scale score data were most representative of the average person in the general population, we found that up to 29% of timed Reading Comprehension scores that would have otherwise been considered average were now classified as impaired relative to higher grade level normative groups. Indeed, when employing the Grade 16 norms, NCE scores decreased by a full standard deviation or more across almost all possible raw scores, relative to Grade 12 norms, and by an average of almost one and a half standard deviations when compared with the pooled norms. The exceptions to this change were scores below the first percentile (raw scores below 7) and the top two or three scores at the highest extreme. Similar but slightly lower decreases were found if norms for Grade 15 were used to interpret identical raw scores from the Reading Comprehension subtest. In other words, as expected with a skewed normative sample made up of mainly higher functioning individuals, scores that would otherwise be considered normal using a more representative comparison group are now said to be impaired relative to those elite few in higher grade levels. If a clinician were employing the older discrepancy method of determining impairment, comparing IQ to achievement, it is clear that large discrepancies would be found if Grade 16 norms were used to interpret obtained raw scores of any individual with an average or above average IQ.

Score changes on the Reading Rate subtest were less extreme regardless of grade level. However, this may be a function of the length of the reading passage. The Technical Manual provides almost no information regarding grade level performance on this measure, nor does it provide any information regarding percent of individuals in each grade who were able to read a specific number of words or the floor and ceiling characteristics of each grade's normative sample. In describing how the Reading Rate measure was developed, the manual refers readers to an obscure paper from 1957 comparing time-limited with rate-limited reading, but the cited paper does not describe a 1-min reading measure.

Similar to the criticisms leveled by Raygor (1978), and Raygor and Flippo (1980) regarding time required to complete the previous version of this test, information provided in the technical manual (Brown et al., 1993b) demonstrated clearly that a large proportion of students in each of the normative grade samples were unable to complete the Comprehension test in 20 min (see Table 7). As may be seen, only 62% of the Grade 16 normative group were able to complete the entire test in the time provided, and only 87% were able to complete even three-quarters of the test under normal time conditions.

The number of students unable to complete most or all of the Reading Comprehension test increased as grade level decreased, such that in Grade 12 and first year college and university samples more than half of students tested could not complete the entire test in the time given. Similarly, a sizeable number were unable to complete 75% of the test in the regular time. Apart from mean scores, no other information regarding the spread of obtained scores for Reading Rate by grade level is provided in the manual.

The Technical Manual notes that the passage used to calculate Reading Rate was chosen such that no student could finish it in 1 min. As a result, it appears that the majority of possible scores are classified as above average without justification as to how this classification was determined. Clearly, the mean score obtained by students at each grade was rated as almost exactly average, so the possible score distribution is positively skewed. For instance, the mean Reading Rate is set at a scaled score of 200 with a SD of 25. Unlike the Reading Comprehension measure, where the highest scale score possible is 250 (two standard deviations above the mean), for Reading Rate the maximum scale score (reading all 601 words in 1 min) is set at 315. This translates into a score that is over four standard deviations above the mean. By contrast, the lowest possible scale score is 136 (2.5 SD below the mean). As such, it is little wonder that more than half of the possible scores listed in the test manual are classified as above average,

Table 7 Percentage who completed all or 75% of the NDRT Reading Comprehension Subtest form G in the allotted time

Grade level sampled	Completed Reading Comprehension test %	Completed 75% of Reading Comprehension test %
Grade 12	47	67
Grade 13 Two-year college	32	55
Grade 14 Two-year college	39	65
Grade 13 Four-year university	44	69
Grade 14 Four-year university	45	76
Grade 15 Four-year university	53	81
Grade 16 Four-year university	62	87

From the Nelson Denny Reading Test Technical Manual (Brown et al., 1993b)

as the score distribution is not symmetrical; the right hand tail of the score distribution is elongated such that it spans over four standard deviations from the mean while the left hand side spans just over two. This information, along with the weak reliability of this measure as reported in the technical manual ($r = 0.68$), makes one question the accuracy of the NDRT Reading Rate score.

Apart from questions regarding the adequacy of the grade-based normative data, one must also consider the ease with which scores on the NDRT could be manipulated. We know that reading impairments are easy to feign and that clinicians are unable to determine when such manipulation has occurred (Harrison, Edwards, & Parker, 2008; Lindstrom, Coleman, Thomassin, Southall, & Lindstrom, 2011). In fact, Lindstrom et al. concluded that students feigning a reading disability produced test score profiles that were "disturbingly sophisticated" (p. 316), easily meeting commonly used diagnostic criteria such as "performing below average on psychoeducational tests". The Reading Rate subtest requires only that the subject self-identify to where they have read in 1 min; slowing down reading speed is one of the main strategies employed by those feigning a reading disorder (Harrison et al., 2008). Hence, it is likely that students actively attempting to feign reading speed impairments could easily produce test scores on the NDRT that would be interpreted as indicating a substantial impairment in timed reading proficiency.

Contrary to what many clinicians may believe, data from the technical manual demonstrates clearly that a substantial proportion of normal, non-disabled students in the NDRT normative samples of each grade were unable to complete the Reading Comprehension subtest in 20 min. In fact, almost 40% of the grade 16 cohort and over half of students between Grades 12 and 15 could not complete the timed Reading Comprehension test in 20 min. Further, 13% of the Grade 16 group were not able to complete even 75% of this test in the normal amount of time provided, a proportion that increased as grade levels decreased. This information has direct implications for clinical interpretation of achieved scores, as clinicians may be tempted to make much of a college student who failed to complete the Reading Comprehension section of the NDRT within the time given. Clearly, this is a common occurrence for many non-disabled students who take the NDRT.

Normative data allow clinicians to determine how abnormal or deviant a score is relative to most other people in the general population. Such scores, in turn, are often used to determine whether a client meets published criteria for a clinical disorder or disability. Standardization samples are supposed to allow the test administrator to compare an individual's performance to that of a normative group. IQ tests, for example, compare raw scores in a given standardization sample to the performance of all individuals of that age. With such age-based norms, test developers try to obtain a broad cross section of same-aged individuals whose abilities are representative of the national population at large. In this way, a test score can answer the question: how similar is this person's performance to most other people in the general population? Further, the ADAAA specifies that this is the standard by which one may determine if an impairment rises to the level of a disability. Our data demonstrate that using postsecondary grade level norms from the NDRT is not appropriate when attempting to determine normative impairment in those who are about to or have already graduated from university.

One is therefore left with an important question: which norms on the NDRT should one employ? If the purpose of the assessment is to determine how an individual is performing relative to a given level of educational attainment (for instance, ranking how well one plays tennis compared with prospective collegiate-level players), then the grade-based scores from the NDRT 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 are not sufficient for the task. Similar to individuals attempting to play collegiate-level tennis, knowing that your otherwise normal abilities do not measure up to the elite competitors you now

face may be important information when making decisions regarding career prospects.

If, however, the purpose is to identify a normative weakness 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 "tennis disabled," one would need to compare her skills to all tennis players to see if she is unable to do what "the majority" of tennis players can achieve. The grade-based postsecondary norms offered by the NDRT fail to provide this normative comparison.

Perhaps because the test manual specifically warns that the NDRT should not be used to diagnose reading disabilities, both the scoring and technical manuals are silent with respect to advising which norms to use when determining normative disability. Comparison of score differences using either Grade 12 or pooled scaled score data as a referent showed that, relatively speaking, little significant difference was obtained in the calculated scores if first year college norms were employed instead. While one could argue that the pooled data likely better represent the performance of most people in the general population, our findings suggest that use of Grade 12 norms as a proxy for the general population likely results in minimal classification changes. By contrast, our results show clearly that use of Grade 15 and especially Grade 16 norms results in substantial score and functioning classification changes when interpreting identical raw reading comprehension scores.

Conclusion

These findings raise questions as to the adequacy of the NDRT when used to determine normative impairment in the timed reading comprehension abilities of young adults in university or postgraduate-level programs. The use of grade norms in particular seems to lead to an over-identification of impairment in such students. Results from this analysis demonstrate clearly that using normative data from the college and university population represents a skewed sample that will, in fact, make otherwise normal individuals appear to be impaired. We would recommend that clinicians employ either Grade 12 end-of-year norms or the scale scores from the pooled normative sample (mean 200, standard deviation of 25) when interpreting the meaning of obtained scores from the NDRT. This latter sample within the NDRT has the largest, most representative group of individuals and likely includes persons from a larger portion of the theoretical normal distribution.

Consistent with the conclusions of both Smith (1998) and Murray-Ward (1998), one should not employ the Reading Rate scores from the NDRT to determine impairment in global speed of reading. Not only does this subtest have a low level of reliability, but it is also a subjective measure that is face valid and therefore easy to feign if one is motivated to demonstrate the need for extra time accommodations.

It remains to be seen whether many of the criticisms identified in this article have been addressed by the newest version of the NDRT, set to be published in late 2018. According to the description given on the test publishers' website (www.proedinc.com), the updated versions (forms I and J) offer age-based norms for adolescents and young adults aged 14 through 25 years that were generated by about 4000 participants stratified to reflect the 2017 US population. However, it appears that the normative data for those over age 18 comes exclusively from postsecondary-level students, which would repeat the same problem regarding a skewed sample of participants as was found in the present study. The new NDRT has also reportedly been found to differentiate students with and without reading problems, although it is unclear whether the former students had a DSM diagnosis of a specific reading disorder.

Finally, and in line with accepted practice standards, even if the NDRT was scored as recommended in this article, clinicians must ensure that the findings from one measure are not interpreted in isolation but in the context of all the other available data and information. In other words, conclusions regarding diagnosis and the need for accommodations should not be based on the findings of only one measure or one test score alone.

Funding 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.

Compliance with Ethical Standards

Conflict of Interest The first author works as a consultant for multiple testing organizations reviewing documentation submitted on behalf of applicants requesting accommodation. The second author declares no conflict of interest.

Informed Consent This article involved no human experimentation or need for informed consent.

Animal Rights No animal studies were carried out by the authors for this article.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

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

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

Bibber v. National Board of Osteopathic Medical Examiners, Inc. (April, 2016). Unites states district court, E. D. Pennsylvania, Civil Action 15 – 4987.

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: Springer.

Brown, J. A., Fishco, V. V., & Hanna, G. (1993a). *Nelson–Denny Reading Test: manual for scoring and interpretation, forms G & H*. Rolling Meadows, IL: Riverside Publishing.

Brown, J. A., Fishco, V. V., & Hanna, G. (1993b). *Nelson–Denny Reading Test: technical report, forms G and H. Manual for scoring and interpretation, forms G & H*. Rolling Meadows, IL: Riverside Publishing.

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*(3), 271–278.

Educational Testing Service, Office of Disability Policy. (2017). *Guidelines for documentation of learning disabilities in adolescents and adults* (4th ed.). Princeton: Author https://www.ets.org/disabilities/documentation/documenting_learning_disabilities/.

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

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.

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*(4), 279–295.

Harrison, A. G., Edwards, M. J., & Parker, K. C. (2008). Identifying students feigning dyslexia: preliminary findings and strategies for detection. *Dyslexia, 14*(3), 228–246.

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).

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

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

Lai, S., & Berkeley, S. (2012). High-stakes test accommodations: research and practice. *Learning Disabilities Quarterly, 35*, 158–169.

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., 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., Coleman, C., Thomassin, K., Southall, C., & Lindstrom, J. (2011). Simulated dyslexia in postsecondary students: description and detection using embedded validity indicators. *The Clinical Neuropsychologist, 25*(2), 302–322.

Lindstrom, W., & Lindstrom, J. (2017). College admissions tests and LD and ADHD documentation guidelines: consistency with emerging legal guidance. *Journal of Disability Policy Studies, 28*(1), 32–42. https://doi.org/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*(4), 611–638. https://doi.org/10.3102/0034654310364063.

Murray-Ward, M. (1998). Test review of the Nelson Denny Reading Test Forms G & H. In J. Impara & B. Plake (Eds.), *The thirteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska Press.

Ofiesh, N., Hughes, C., & Scott, S. (2004). Extended test time and postsecondary students with learning disabilities: a model for decision making. *Learning Disabilities Research and Practice, 19*, 57–70.

Phillips, S. E. (1994). High-stakes testing accommodations: validity versus disabled rights. *Applied Measurement in Education, 7*(2), 93–120. https://doi.org/10.1207/s15324818ame0702_1.

Raygor, A. L. (1978). Nelson-Denny Reading Test, Forms C and D. In K. Burros (Ed.), *Eighth mental measurements yearbook*. Highland Park, New Jersey: The Gryphon Press.

Raygor, A. L. & Flippo, R. F. (1980). Varieties of comprehension measures: a comparison of intercorrelations among several reading tests. Arlington, Virginia: *ERIC Document Reproduction Service.* (ERIC Document Reproduction Service No. ED 193 485).

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*(1), 72–80. https://doi.org/10.1177/0829573512437171.

Smith, D. (1998). Test review of the Nelson Denny Reading Test Forms G & H. In J. Impara & B. Plake (Eds.), *The thirteenth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements, University of Nebraska Press.

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.

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: Lawrence Erlbaum.

U.S. Census Bureau retrieved October 2, 2017 https://www2.census.gov/programs-surveys/demo/tables/educational-attainment/time-series/p20-476/tab18.pdf

Van Meter, B. J., & Herrmann, B. A. (1986). Use and misuse of the Nelson-Denny Reading Test. *Community College Review, 14*(3), 25–31.