Journal of Psychoeducational Assessment
2015, Vol. 33(4) 299–311
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0734282915573723
jpa.sagepub.com
SAGE

Implications for Educational Classification and Psychological Diagnoses Using the Wechsler Adult Intelligence Scale–Fourth Edition With Canadian Versus American Norms

Allyson G. Harrison[1], Alana Holmes[2], Robert Silvestri[2], and Irene T. Armstrong[1]

Abstract

Building on a recent work of Harrison, Armstrong, Harrison, Iverson and Lange which suggested that Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV) scores might systematically overestimate the severity of intellectual impairments if Canadian norms are used, the present study examined differences between Canadian and American derived WAIS-IV scores from 861 postsecondary students attending school across the province of Ontario, Canada. This broader data set confirmed a trend whereby individuals' raw scores systematically produced lower standardized scores through the use of Canadian as opposed to American norms. The differences do not appear to be due to cultural, educational, or population differences, as participants acted as their own controls. The ramifications of utilizing the different norms were examined with regard to psychoeducational assessments and educational placement decisions particularly with respect to the diagnoses of Learning Disability and Intellectual Disability.

Keywords

intelligence, normative data, Canadian, adult, WAIS

The value of almost all psychological tests is in the extent to which a score obtained by a specific individual may be compared with his or her peer group and ranked relative to the average score. The need for appropriate normative data has been exemplified in the measurement of general intelligence. Indeed, the Wechsler family of intelligence tests (e.g., Wechsler Intelligence Scale for Children [WISC], Wechsler Adult Intelligence Scale [WAIS], Wechsler, 1949, 1991, 2004a) are some of the most widely administered in the world (Kaplan & Saccuzzo, 2005; Plante, 2010), and educational psychologists rely on these tests to estimate the overall and specific intellectual

[1] Queens University, Kingston, Ontario, Canada
[2] Cambrian College, Sudbury, Ontario, Canada

Corresponding Author:
Allyson G. Harrison, Regional Assessment & Resource Centre, Queen's University, Mackintosh-Corry Hall, B100, 68 University Avenue, Kingston, Ontario K7L 3N6, Canada.
Email: harrisna@queensu.ca

skills of a given individual relative to his or her same-aged peers. In this way, psychologists are able to determine whether a given individual is in some way exceptional. For instance, does the student have adequate intellectual ability necessary for learning, one of the criteria that must be demonstrated before making the diagnosis of a Learning Disability (LD; for example, Flanagan, Alfonso, & Mascolo, 2011; Ontario Ministry of Education, 2014); should they be referred for gifted programming (Gross, 2004; Johnsen, 2004; Pfeiffer, 2012); would they qualify for identification as a student with an Intellectual Disability (ID; Bergeron, Floyd, & Shands, 2008; McDermott, Watkins, & Rhoad, 2014); or should one take intellectual impairments into account when treating a mental health condition (e.g., Kamphaus, Worrell, & Harrison, 2005; McDermott et al., 2014). All such diagnostic decisions, however, rest on the assumption that the test utilized accurately ranks the client's ability relative to his or her peer group.

Until recently, the general intellectual ability of Canadian individuals was calculated by comparing the obtained test scores for a single individual with normative data collected from American individuals. Some Canadian psychologists (e.g., Beal, 1988) expressed concern that comparison with Americans might not be the most appropriate way to determine the intellectual ability of Canadians, prompting the Psychological Corporation (now Pearson) to evaluate whether data from both populations were equivalent (e.g., Wechsler, 1996, 2001, 2004a, 2004b; Wechsler & Naglieri, 2006).

Initially, distinct normative data were obtained in the updated version of the Wechsler Intelligence Scale for Children–Third Edition (WISC-III; Wechsler, 1996). Indeed, validation data collected during the updated norming of this test found significant differences between performance of Canadian and American children, with Canadian children obtaining consistently higher raw scores on each subtest of the test battery than their American counterparts. As a result, the Psychological Corporation published separate normative data for American and Canadian clients.

Similar differences were identified when data were collected for the Wechsler Adult Intellectual Scale–Third Edition (WAIS-III; Wechsler, 2001) and also the Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV; Wechsler, 2008a, 2008b), with Canadian adults typically obtaining higher raw scores on each of the WAIS subtests relative to their American counterparts, although such differences were less apparent in individuals above age 65 (Wechsler, 2001, 2008a). This resulted in the publication of separate normative data for Canadian and American adults.

It was posited that the differences in Canadian and American scores were the results of varying population demographics between countries particularly, ethnicity and educational achievement factors (WAIS-IV Canadian technical manual, Wechsler, 2008a). Preliminary support for this proposition was obtained during the Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV) standardization wherein all composite score differences between the two countries were mitigated when respective norming samples were matched on key social, economic, and educational demographics (Wechsler, 2004b).

Although it is not known exactly why these differences occur, it is clear that applying different normative data to analyze obtained scores does result in clinically different conclusions for a sizable number of adults. For example, Iverson, Lange, and Viljoen (2006) found that clinically different conclusions could be obtained when comparing Canadian and American data on 13% to 21% of WAIS-III individual subtest or index scores for a sample of psychiatry and neuropsychiatry inpatients. Directionality of these score changes was consistent in that lower Canadian scores relative to American scores were found with significant variation on Full Scale IQ (FSIQ) and Performance IQ. These authors conclude that "significantly lower scores on all IQ, indices, and subtest scores will be calculated when using the Canadian versus the American norms" (Iverson et al., 2006, p. 351). Given that a higher number of patients returned more subtest scores in the impaired range when using Canadian norms, the authors note that clinicians might conclude greater impairment in intellectual abilities than if the American norms were used. Because of the significantly large difference in scores, these authors advise clinicians not to mix Canadian-derived

IQ scores with American-derived achievement tests or vice versa, and to ensure that they had good reasons for choosing one set of normative data over the other prior to calculating the score of an individual client. Similar findings were reported by Beal, Dumont, Cruse, and Branche (1996) when examining the performance of children on the WISC-III.

Recently, the adult version of the Wechsler Intelligence Scale was revised and updated, requiring collection of new normative data in both America (WAIS-IV; Wechsler, 2008b) and Canada (Wechsler, 2008a). In America, the WAIS-IV was standardized on a *normative sample* of 2,200 individuals between the ages of 16 and 90, stratified into 13 age groups. An extension of the standardization was then conducted with 688 Canadians in the same stratified age range. Inferential norming was used for the development of both normative samples, with the raw Canadian data matched to the distribution produced in the American standardization. With inferential norming, "statistics (such as mean or standard deviation) can be plotted and fixed to a polynomial regression curve to estimate trends across inferential demographic variables such as age and grade" (Zhu & Chen, 2011, p. 2). As such, this technique allows the user to estimate the characteristic of any variable for each age group without requiring a large sample size. Ideally, individually administered tests such as the WAIS would have very large samples for each age, education, and sex group; the total size of the sample collected for the Canadian norms is therefore smaller than is considered ideal in a purely statistical sense.

Inferential norms are said to provide a more accurate estimation of population parameters such as means and standard deviations because they are based on an equation that results from using the data for all demographic groups, rather than data from only one group, for a particular table (Zachary & Gorsuch, 1985). Thus, information about the effects of age, education, and sex on WAIS standardized scores derived from the entire sample of 688 participants is used to determine the normative performance for each age, education, and sex group (i.e., normative table).

Previous research has indicated that the raw scores obtained in both the Canadian and American normative samples demonstrate similar general construct validity (i.e., raw scores from both data sets appear to conform to a modified four-factor model with similar factor loadings across the two standardization samples; Bowden, Saklofske, & Weiss, 2011a, 2011b). Furthermore, the American norms have been shown to result in consistent constructs between normative and clinical samples (Weiss, Keith, Zhu, & Chen, 2013) although not uncontroversially (see Canivey & Kush, 2013; Grégoire, 2013). However, it is unclear whether the standardized scores computed from each data set are equivalent in their distribution and score classification.

Recently, Harrison, Armstrong, Harrison, Lange, and Iverson (2014) investigated the difference between Canadian and American normative scores for the WAIS-IV in a sample of students (n = 432) from Southern Ontario, Canada. They found score differences greater than those identified by Iverson et al. (2006) and concluded that neuropsychologists would likely obtain very different results depending on the normative data used. They pointed to the problems this could create in neuropsychological evaluations and called for additional research to investigate the impact that such differences would have on assessment and diagnosis of other types of intellectual disorders.

The present study sought to expand on the work of Harrison et al. (2014) by investigating the effect of using different normative data when calculating scores on the WAIS-IV for postsecondary students undergoing psychoeducational assessments across all regions of Ontario, Canada. Such information is particularly important because norming differences may influence score interpretation and diagnosis (Iverson et al., 2006; Ryan & Schnakenberg-Ott, 2003), especially when IQ cut scores are employed in specific diagnostic classification schemes [e.g., identification of LD (in Ontario, Canada students must have average FSIQ or General Abilities Index (GAI; Ontario Ministry of Education, 2014); identification of gifted individuals (e.g., Gross, 2004; Johnsen, 2004); or identification of ID (Harrison & Holmes, 2014)].

It was hypothesized that statistically significant differences would be found on all IQ, index, and subtest scaled scores, with Canadian normative scores being lower than American normative scores. It was also hypothesized that such differences would be clinically meaningful (i.e., 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

Participants

Participants in this ethics-approved study were taken retrospectively from the databases of two regional assessment centers that serve the entire province of Ontario. Approximately half of these protocols (432/861) were reported on previously by Harrison et al. (2014), reflecting data collected in the southern half of the province. The current sample now includes data from postsecondary schools in the northern half of the province as well as additional cases obtained since submission of the previous article. All participants in the database had been referred for a psychoeducational or neuropsychological evaluation of a previously identified or suspected learning and/or attention problems, and had given consent to have their data used in future research studies. Students referred to these assessment services had either been accepted into or were currently studying at a postsecondary institution somewhere in the province of Ontario. Thirty-seven percent were enrolled in university, and 63% were enrolled in college programs; all had met the normal academic entrance requirements for their respective programs and were not enrolled in any modified or upgrading courses. The sample included recent high school graduates transitioning directly into postsecondary studies and mature students who often had completed some form of scholastic upgrading; it also included a small proportion of graduate students referred due to recently reported academic difficulties.

Although many of the referred students did receive a diagnosis of a specific LD or an Attention Deficit Hyperactivity Disorder (53.5% in total), many others (35.5%) were found to have learning or attention problems due to other causes such as generalized anxiety, depression, borderline intellectual functioning, obsessive-compulsive personality disorder, perfectionism, or weak academic background, and the rest (11.0%) received no diagnosis (i.e., they were normal). The diagnostic outcomes and intellectual levels of functioning for the present sample can thus be summarized as variable in nature.

In total, 861 complete WAIS-IV protocols were available for review in this study. The sample was 42.4% male. Age of participants ranged from 16 to 63 years ($M = 23.7$, $SD = 8$). Although the breakdown of ethnicity was not coded, the majority of students were Caucasian.

Procedure

Each participant was administered the WAIS-IV as part of a comprehensive psychoeducational assessment that also included measures of performance validity (see Larrabee, 2012, for a discussion of performance validity). The raw scores from each of the WAIS-IV subtests were then entered into the computer scoring program (PsychCorpCenter–II). This program allows for raw score interpretation using either Canadian or American norms. Resulting scores from the application of each set of normative data were then entered into a database. Protocols with score differences greater than half of a standard deviation were rescored to ensure accurate calculation of index and scaled scores.

Results

Descriptive statistics (i.e., means, standard deviation, correlation coefficients) and effect sizes regarding composite, index, and subtest scores using both Canadian and American normative

Table 1. Descriptive Statistics, Correlations, Mean Comparisons, and Effect Sizes.

	American norms		Canadian norms				
Score	*M*	*SD*	*M*	*SD*	*p*	*d*	*r*
Composite Scores							
FSIQ	95.5	12.9	88.1	14.4	<.001	.54	.99
General Ability Index (*n* = 800/828)	98.9	14.7	92.5	16.0	<.001	.42	.99
Index Scores							
Verbal Comprehension	97.9	15.1	91.8	16.3	<.001	.39	.99
Perceptual Reasoning	99.9	14.1	94.5	15.9	<.001	.36	.99
Working Memory	90.5	12.8	83.5	13.8	<.001	.53	.99
Processing Speed	95.2	12.9	90.4	14.1	<.001	.36	.99
Subtest Scores							
Verbal Subtests							
Vocabulary	9.9	3.1	8.7	3.3	<.001	.37	.99
Similarities	9.7	3.0	8.5	3.3	<.001	.38	.98
Information	9.2	3.1	8.5	3.3	<.001	.22	.99
Arithmetic	8.2	2.7	7.4	2.7	<.001	.30	.99
Digit Span	8.4	2.5	7.1	2.7	<.001	.50	.98
Performance Subtests							
Block Design	9.8	3.0	8.9	3.2	<.001	.29	.99
Matrix Reasoning	9.8	2.9	9.1	3.2	<.001	.23	.99
Visual Puzzles	10.5	2.9	9.4	3.1	<.001	.37	.99
Symbol Search	9.3	2.8	8.5	3.0	<.001	.28	.99
Coding	8.9	2.5	8.2	2.6	<.001	.27	.98

Note. n = 861-852 for all variables unless shown in parentheses beside variable name. Within parentheses, American *n*/Canadian *n. r* is a correlation coefficient. FSIQ = Full Scale IQ.

data may be found in Table 1. As may be seen, the mean FSIQ using American norms is average (95.5), whereas it is low average using Canadian norms (88.1). The correlation between the two normative systems is high, with correlation coefficients ranging from .98 to .99. Using paired-sample *t* tests and applying a Bonferroni correction, significantly lower ($p < .001$) scores were found on all composite, index, and subtest scores when using Canadian as opposed to American norms. Furthermore, medium effect sizes were found when examining this score difference for FSIQ, GAI, Working Memory Index (WMI), and Digit Span (*d* range = .42-.54). Effect sizes for the remaining index and subtest scores ranged from small (*d* = .22) to medium (*d* = .39).

Difference scores between Canadian- and American-generated composite, index, and subtest scores were calculated by subtracting corresponding values between the two normative systems. As shown in Table 2, the majority of the sample obtained higher index, composite, and subtest scores when American norms were applied (FSIQ range = −1 to 13; GAI range = −1 to 12; index scores range = −5 to 16; and subtest scores range = −2 to 4). Apart from scores on the Processing Speed Index (PSI), no individual below age 35 obtained a higher score on FSIQ or an index score when Canadian norms were used.

The percentage agreement between the two normative systems was then calculated using two criteria: (a) percent within 1/3 of a *SD* (i.e., the standard error of measurement for this test, ±5 points for composite and index scores; 1 point for subtest scores); (b) percent within the same ability classification level (ranging from *extremely low* to *very superior*); or both (percent within 1/3 of a *SD or* within the same ability classification). The highest rate of agreement using both criteria across the composite and index scores was found for the Perceptual Reasoning Index

Table 2. Mean Differences and Percentage Agreement Among Normative Systems of the WAIS-IV: Composite, Index, and Subtest Scores (*n* = 857-852).

	M difference (*SD*)	% within ±5 points	% within same classification	% within ±5 points or same classification
Composite Scores				
FSIQ	7.5 (2.3)	19.4	38.7	44.9
General Ability Index	6.4 (2.2)	35.0	47.4	56.4
Index Scores				
Verbal Comprehension	6.0 (2.4)	38.9	53.6	64.5
Perceptual Reasoning	5.4 (2.6)	54.2	63.6	73.6
Working Memory	7.0 (2.4)	23.5	40.8	45.0
Processing Speed	4.9 (2.6)	53.0	58.1	70.8
	M difference (*SD*)	% within ±1 point	% within same classification	% within ±1 point or same classification
Verbal Subtests				
Vocabulary	1.2 (0.6)	71.0	57.9	74.0
Similarities	1.2 (0.6)	69.6	60.4	73.9
Information	0.7 (0.5)	99.5	79.0	99.6
Arithmetic	0.9 (0.4)	97.7	67.1	97.5
Digit Span	1.2 (0.6)	69.8	52.2	74.0
Performance Subtests				
Block Design	0.9 (0.5)	90.8	64.8	90.9
Matrix Reasoning	0.7 (0.6)	93.4	76.4	93.8
Visual Puzzles	1.1 (0.6)	78.9	61.1	80.4
Symbol Search	0.8 (0.5)	94.5	78.0	94.8
Coding	0.7 (0.5)	99.4	75.8	99.4

Note. Difference is American norms minus Canadian norms, positive values indicate greater American norms. WAIS-IV = Wechsler Adult Intelligence Scale–Fourth Edition; FSIQ = Full Scale IQ.

(PRI; 73.6%) and PSI (70.8%). Lower rates of agreement were found for the remaining composite and index scores (ranging from 44.9% for FSIQ to 64.5% for Verbal Comprehension). Six of the 10 subtest scores (with the exception of Vocabulary, Similarities, Digit Span, and Visual Puzzles) had high rates of agreement, with Information showing the highest agreement (99.6%).

Of interest was the percentage of individuals who would be classified as having a FSIQ below the 10th percentile or who would fall within the IQ range required for diagnosis of ID (e.g., 70 ± 5) when both normative systems were applied to the same raw scores. Using American norms, 13.1% had an IQ of 80 or less, and 4.2% had an IQ of 75 or less. By contrast, when using Canadian norms, 32.3% had an IQ of 80 or less, and 21.2% had an IQ of 75 or less. Most notably, only 0.7% (2 individuals) obtained a FSIQ of 70 or less using American norms, whereas 9.7% had IQ scores this low when Canadian norms were used. At the other end of the spectrum, 1.4% of the students had FSIQ scores of 130 or more (gifted) when American norms were used, whereas only 0.3% were this high using Canadian norms.

When defining average as a FSIQ in the range of 90 to 109 and using American norms, 51.6% of the sample was classified as average. The percentage dropped substantially, however, with the evocation of the Canadian norms with only 34.2% of the sample receiving the label of average.

Score differences were next examined based on different levels of intellectual classification to determine whether systematic differences occurred across all levels of IQ or whether they were more pronounced in certain areas of the score distribution. As may be seen in Table 3, the

Table 3. Percentages of Students With Clinically Differing Scores Between Normative Systems of the WAIS-IV as a Function of American FSIQ Level.

	American FSIQ level						
	<70 (*n* = 2)	70-79 (*n* = 89)	80-89 (*n* = 203)	90-109 (*n* = 443)	110-119 (*n* = 80)	120-130 (*n* = 30)	>130 (*n* = 12)
Composite Scores							
FSIQ	0.0	68.7	85.1	44.6	27.5	23.3	33.3
General Ability Index	50.0	64.6	69.7	32.6	34.2	22.2	0.0
Index Scores							
Verbal Comprehension	50.0	61.8	53.2	26.5	25.0	13.3	0.0
Perceptual Reasoning	50.0	77.5	43.3	13.3	7.5	10.0	8.3
Working Memory	0.0	69.7	73.4	51.6	33.7	13.3	16.7
Processing Speed	0.0	57.3	38.1	24.2	13.7	13.3	0.0
Verbal Subtests							
Vocabulary	100.0	58.4	39.4	15.2	17.5	20.0	16.7
Similarities	0.0	42.7	30.5	14.9	13.5	3.3	0.0
Information	100.0	0.0	1.5	0.0	0.0	0.0	0.0
Arithmetic	100.0	5.6	2.5	2.3	0.0	0.0	0.0
Digit Span	50.0	34.8	34.7	23.6	17.5	6.7	0.0
Performance Subtests							
Block Design	50.0	28.1	16.3	4.3	0.0	0.0	0.0
Matrix Reasoning	100.0	10.1	11.9	4.5	0.0	0.0	0.0
Visual Puzzles	0.0	48.3	25.6	13.7	13.5	3.3	0.0
Symbol Search	100.0	11.2	6.5	3.9	2.5	6.7	0.0
Coding	100.0	1.1	1.0	0.5	0.0	0.0	0.0

Note. WAIS-IV = Wechsler Adult Intelligence Scale–Fourth Edition; FSIQ = Full Scale IQ.

percentage of individuals whose scores would be interpreted as clinically similar (i.e., within the same classification level) was consistently smaller as overall FSIQ decreased. Indeed, the mean difference in scores between the two systems is almost a full *SD* at the lowest IQ classification, whereas it is about one third of a *SD* (4.8) in the highest classification range. This same trend was evident for composite, index, and subtest scores, with larger score differences occurring as FSIQ declined. Apart from two students with American-derived FSIQ scores of 70 or less, individuals in this study had a greater chance of being classified in a lower IQ range as their calculated FSIQ decreased.

We also examined the effect across age cohort on score differences between the two normative systems. Table 4 shows the percentage of individuals who change composite, index, and subtest classifications according to WAIS-IV age cohort. As may be seen, significantly larger differences (about half a *SD*) were found for composite and index scores in those below age 35. Notably, more than half of the individuals below age 35 change FSIQ classification, and just less than half change GAI classification.

Discussion

The FSIQ and then the GAI are the IQ scores most frequently recommended for deployment within school board systems and other agencies charged with the task of classifying individuals who require specialized educational supports or services. Classification is typically required to access the most appropriate forms of learning supports but also to ensure that limited and

Table 4. Percentages of Students With Clinically Differing Scores Between Normative Systems by Age Group.

	Age in years							
	16-17 (*n* = 32)	18-19 (*n* = 318)	20-24 (*n* = 264)	25-29 (*n* = 111)	30-34 (*n* = 45)	35-44 (*n* = 52)	45-54 (*n* = 26)	55-64 (*n* = 9)
	%	%	%	%	%	%	%	%
Composite Scores								
FSIQ	53.1	61.5	51.9	56.8	53.3	39.2	38.5	33.3
GAI	41.4	47.3	43.7	49.0	41.5	34.7	11.5	11.1
Index Scores								
Verbal Comp.	31.2	41.6	36.1	29.7	42.2	17.3	11.5	33.3
Perceptual Reasoning	31.2	24.9	27.0	37.8	22.2	23.1	7.7	0.0
Working Memory	40.6	56.5	61.5	62.2	51.1	30.8	19.2	33.3
Processing Speed	34.4	35.0	26.7	25.2	28.9	21.6	15.4	0.0
Verbal Subtests								
Vocabulary	21.9	35.3	21.7	18.9	26.7	13.5	23.1	0.0
Similarities	15.6	13.8	26.9	29.7	44.4	0.0	3.8	44.4
Information	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0
Arithmetic	3.2	1.3	0.0	13.5	0.0	1.9	0.0	0.0
Digit Span	0.0	37.7	25.9	29.7	0.0	1.9	0.0	0.0
Performance Subtests								
Block Design	12.5	5.7	11.4	16.2	4.4	9.6	0.0	0.0
Matrix Reasoning	0.0	5.1	6.4	12.6	11.1	1.9	0.0	0.0
Visual Puzzles	16.1	16.5	26.4	26.1	4.4	17.3	0.0	0.0
Symbol Search	6.5	12.6	0.0	0.9	0.0	0.0	0.0	0.0
Coding	0.0	0.6	0.8	0.0	0.0	2.0	0.0	0.0

Note. FSIQ = Full Scale IQ; GAI = General Ability Index; Verbal Comp. = Verbal Comprehension.

valuable services are directed to those most in need of them. The present data set comparing American and Canadian normative scores (derived from the same raw score for an individual client) shows that FSIQ and GAI scores consistently and significantly decrease when Canadian norms are utilized; only 3 of 861 individuals (all in the 45-54 age range) obtained a higher FSIQ score when Canadian norms were used, and only 2 individuals (above age 35) obtained higher GAI scores when using these norms. This finding is similar to that noted by Harrison et al. (2014) whose work showed that using Canadian norms resulted in a higher FSIQ in only one individual from their sample. This downward shift is also reflected in the fact that the FSIQ scores of 61.3% of students and the GAI scores of 52.6% of students dropped a classification level when their raw scores were scaled using Canadian as opposed to American norms. Even more to the point, a sizable number of students' scores dropped in a clinically meaningful manner when moving from the American norms to Canadian norms. That is to say, 45% of the students classified as average, and 85% of the students classified as below average by American norms dropped classification categories when their IQ scores were computed using Canadian norms. In both instances, this downward shift has implications for the rendering of diagnoses: LD in the first instance and ID in the latter, wherein students formerly eligible for a diagnosis of LD would no longer be said to have otherwise average thinking and reasoning abilities (something required by the Ontario Ministry of Education, 2014, in their definition of a LD), and students previously without a diagnostic label might now be considered for a diagnosis of ID. In other words, the prevalence of educational disorders stands to wax and wane with the selection of norms.

Because issues such as level of student engagement and test measurement error can affect performance during cognitive testing, using test scores in the absence of collaborative data is not recommended for the formulation of diagnoses that have level of intellectual functioning as one of their cardinal features. Indeed, respected diagnostic codebooks such as the International Classification of Diseases (ICD, World Health Organization, 2010) and the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; *DSM-5*; American Psychiatric Association, 2013) acknowledge this within their criteria when they refer to the need for historical evidence to be considered alongside test scores as means to speaking to the ecological validity of test scores. That said, research shows that IQ scores are still "routinely used to differentially classify mental disability" (McDermott et al., 2014, p. 207), and students in school boards across Ontario, Canada, are often assigned an exceptionality classification of "Mild Intellectual Disability" on the basis of cognitive test scores alone (Harrison & Holmes, 2014). Assuming that intelligence is normally distributed, one would predict that only about 10% of the population would have a FSIQ of 80 or less, and that only about 5% would have a FSIQ of 75 or less. Although only 13.1% of the students in the current sample returned a FSIQ of 80 or less using American norms, 32.3% had a score this low when Canadian norms were used. The increase is also striking for FSIQs of 75 or less: 4.2% of scores fall at or below this cutoff when using American norms compared with 21.2% when using Canadian norms. Although not as large at the other end of the IQ spectrum, one also finds differences in classification of gifted students: 1.4% of IQ scores had a FSIQ of more than 130 using American norms, whereas only 0.3% were this high using Canadian norms. Clearly, the scores obtained using Canadian norms are significantly different than what would be predicted from the normal distribution or bell curve. Given that IQ scores are often used in Canada to identify abnormalities in functioning, such scoring differences could have broad implications for the number of students in Canada who would qualify as having a disability. For instance, it seems impossible to believe that 21% of students accepted into postsecondary degree or diploma programs in Ontario have FSIQ scores in the intellectually deficient range.

Similar to the previous finding of Harrison et al. (2014), the present findings also reveal that the indices that comprise the FSIQ and GAI are most variable when they fall at the lower end of the normal curve (American FSIQ of 90 or below, see Table 3) or are applied to younger postsecondary students (ages 16-34, see Table 4). This highlights a bias with respect to individuals whose intellectual functioning is on the lower side of average, wherein 100 is taken to represent average. However, individuals with scores in this range are those who are most likely to present for assessment due to their poor progress or learning challenges within the academic environment. In other words, the downward classification drift is most likely to affect those with learning needs as opposed to those without such challenges.

Our study identified that larger discrepancies between American and Canadian-derived scores were found in those individuals whose IQs were below average. It has been argued (Pearson, 2014) that the reason for this finding is that the *SD* in the Canadian normative sample was smaller than that obtained in the American normative study. Hence, when the raw scores from the Canadian sample are made to "fit" the American-derived normal curve, which has a larger *SD*, those individuals who are below average are pulled down on the bell curve, meaning that their raw scores subsequently correspond to lower scaled scores when matched to the American-derived data. If true, then this should also mean that those individuals whose scores are above average should enjoy the opposite, that is, an elevation in their scores when the Canadian norms are applied. Such, however, was not found to be the case. Indeed, across all levels of intelligence, the trend is for students to obtain lower scores on all subtest, index, and IQ scores when the Canadian norms are applied, but with those below the average range experiencing a greater drop in score relative to those whose IQ or index scores were average or higher. This trend was also most pronounced in the younger age groups.

How is it that selecting Canadian over American norms so markedly lowers the standard scores generated from the identical raw scores? One possible explanation is that more extreme scores occur because the Canadian normative sample is smaller than the American (cf. Kahneman, 2011); another explanation is that sampling bias has occurred whereby the sample includes individuals with higher than typical levels of education, more urban as opposed to rural participants and underrepresents aboriginal groups, especially in the younger age groups (see Harrison et al., 2014 for a more thorough discussion). If fewer lower IQ individuals were actually sampled in the lower age ranges originally, then any regression-based scores derived would tend to perform more poorly in extreme score ranges. This could explain why 48.9% of postsecondary-aged students in the present sample were said to have FSIQ scores of 85 or less using Canadian norms, whereas if IQ scores are distributed normally, one would predict that only about 16% of the population should have scores in this range.

One cannot explain this difference simply by saying it is due to the mature students in the sample who completed academic upgrading, as the score differences were most prominent in the youngest cohorts. It is difficult to explain these findings simply as a function of disability status, as all participants were deemed otherwise qualified by these postsecondary institutions (i.e., they had met normal academic requirements for entry into regular postsecondary programs). Furthermore, in Ontario, a diagnosis of LD is given only to students with otherwise normal thinking and reasoning skills, and so students with such a priori diagnosis would have had otherwise average full scale or general abilities scores when tested previously. Performance exaggeration seems an unlikely cause for the findings, as the students' scores declined only when Canadian norms were applied. Finally, although no one would argue that a subset of disabled students might be functioning below average, it is difficult to believe that almost half of these postsecondary students would fall in this IQ range given that they had graduated from high school with marks high enough to qualify for acceptance into bona fide postsecondary programs. Whatever the cause, our data suggest that one must question both the representativeness of the Canadian normative sample in the younger age ranges and the accuracy of the scores derived when these norms are applied.

Differences in derived composite, index, and subtest scores could also have serious implications in both forensic and memory disorders settings. For instance, an individual whose premorbid functioning is being estimated after an accident may appear to have suffered minimal cognitive decline relative to the average if Canadian norms were used to score the WAIS, as these scores may underestimate the person's functioning on tests known to be less sensitive to the effects of brain injury such as Vocabulary. Of more concern, clinicians might also choose to use Canadian norms to calculate IQ scores if they are trying to advocate that a client suffered a serious brain injury, even when the American-based scores indicate that the person is functioning normally. Conversely, cognitive declines associated with Alzheimer's or other dementias might be incorrectly suspected if an otherwise average individual is assessed using the Canadian version of the WAIS.

This study had a few limitations. First, although the sample size was larger than that obtained in the Canadian norming of the WAIS-IV, it was not obtained in a random manner. Students were referred either to verify the presence of a previously identified learning or attention problem, or to investigate whether such a disability was present. Even so, the IQs in our sample were distributed normally when derived using American norms, and a number of the individuals in the sample were not experiencing any academic impairments. Given the difficulties in obtaining a randomly chosen sample of this magnitude, however, it seems unlikely that a random sample of this size could be obtained by anyone other than a large test publisher. It is also true that some cell sizes in our sample were quite small (e.g., FSIQ below 70; individuals above age 54), and so the numbers obtained in those cells are likely unstable. Nevertheless, the robust findings demonstrated in cells with sufficient numbers support our conclusion that the Canadian norms systematically result in lower scaled scores, and that those with IQs below average or below age 35 are most affected by this scoring difference. Finally, we did not track ethnic background in our database. However, each participant acted as his or her own control, so it seems difficult to suggest that cultural factors are responsible for our findings given that they should affect the obtained scores of each individual to an equal extent.

Overall, our findings suggest a need to examine more carefully the accuracy and applicability of the WAIS-IV Canadian norms when interpreting raw test data obtained from Canadian adults. Using these norms appears to increase the number of young adults identified as intellectually impaired and could decrease the number who qualify for gifted programming or a diagnosis of LD. Until more research is conducted, we strongly recommend that clinicians not use Canadian norms to determine intellectual impairment or disability status. Converting raw scores into Canadian standard scores, as opposed to using American norms, systematically lowers the scores of postsecondary students below the age of 35, as the drop in FSIQ was higher for this group than for older adults. Although we cannot know which derived scores most accurately reflect the intellectual abilities of young Canadian adults, it certainly seems implausible that almost half of postsecondary students have FSIQ scores below the 16th percentile, calling into question the accuracy of all other derived WAIS-IV Canadian scores in the classification of cognitive abilities.

Authors' Note

The views expressed in this article do not necessarily reflect those of the Ministry of Training, Colleges, and Universities.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported in part by funding from the Ministry of Training, Colleges, and Universities (MTCU) of Ontario.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Beal, A. L. (1988). Canadian content in the WISC-R: Bias or jingoism. *Canadian Journal of Behavioral Science*, *20*, 154-166. doi:10.1037/h0079921

Beal, A. L., Dumont, R., Cruse, C. L., & Branche, A. H. (1996). Practical implications of differences between the American and Canadian norms for WISC-III and a short form for children with learning disabilities. *Canadian Journal of School Psychology*, *12*(1), 7-14. doi:10.1177/082957359601200103

Bergeron, R., Floyd, R. G., & Shands, E. I. (2008). State eligibility guidelines for mental retardation: An update and consideration of part scores and unreliability of IQs. *Education and Training in Developmental Disabilities*, *41*, 123-131.

Bowden, S., Saklofske, D. H., & Weiss, L. G. (2011a). Augmenting the core battery with supplementary subtests: Wechsler Adult Intelligence Scale–IV measurement invariance across the United States and Canada. *Assessment*, *18*, 133-140.

Bowden, S., Saklofske, D. H., & Weiss, L. G. (2011b). Invariance of the measurement model underlying the Wechsler Adult Intelligence Scale–IV in the United States and Canada. *Educational and Psychological Measurement*, *71*, 186-199.

Canivey, G. L., & Kush, J. C. (2013). WAIS-IV and WISC-IV structural validity: Alternate methods, alternate results. Commentary on Weiss et al. (2013a) and Weiss et al. (2013b). *Journal of Psychoeducational Assessment*, *31*(2), 157-169.

Flanagan, D. P., Alfonso, V. C., & Mascolo, J. T. (2011). A CHC-based operational definition of SLD: Integrating multiple data sources and multiple data gathering methods. In D. P. Flanagan & V. C. Alfonso (Eds.), *Essentials of specific learning disability identification* (pp. 233-298). New York, NY: John Wiley.

Grégoire, J. (2013). Measuring components of intelligence: Mission impossible? *Journal of Psychoeducational Assessment*, *31*(2), 138-147.

Gross, U. M. (2004). *Exceptionally gifted children* (2nd ed.). London, England: Routledge Falmer. (Original work published 1993)

Harrison, A. G., Armstrong, I. T., Harrison, L. E., Lange, R. T., & Iverson, G. L. (2014). Comparing Canadian and American Normative scores on the Wechsler Adult Intelligence Scale–Fourth Edition. *Archives of Clinical Neuropsychology*, *29*, 737-746.

Harrison, A. G., & Holmes, A. (2014). Mild intellectual disability at the postsecondary level: Results of a survey of disability service offices. *Exceptionality Education International*, *23*(1), 22-39.

Iverson, G. L., Lange, R. T., & Viljoen, H. (2006). Comparing the Canadian and American WAIS-III normative systems in inpatient neuropsychiatry and forensic psychiatry. *Canadian Journal of Behavioural Science*, *38*, 348-353.

Johnsen, S. K. (2004). *Identifying gifted students: A practical guide*. Waco, TX: Prufrock Press.

Kahneman, D. (2011). *Thinking fast and slow*. Toronto, Ontario, Canada: Macmillan.

Kamphaus, R., Worrell, F. C., & Harrison, P. (2005). Principles for evaluation and eligibility determination for specific learning disabilities: A report of the ad hoc committee of division 16. *The School Psychologist*, *59*, 157-159.

Kaplan, R. M., & Saccuzzo, D. P. (2005). *Psychological testing: Principles, applications, and issues* (6th ed.). Belmont, CA: Thompson Wadsworth.

Larrabee, G. J. (2012). Performance validity and symptom validity in neuropsychological assessment. *Journal of the International Neuropsychological Society*, *18*, 625-630.

McDermott, P., Watkins, M., & Rhoad, A. (2014). Whose IQ is it?—Assessor bias variance in high-stakes psychological assessment. *Assessment*, *26*, 207-214. doi:10.1037/a0034832

Ontario Ministry of Education. (2014). *Identification of and program planning for students with learning disabilities* (Policy/Program memorandum No. 8). Toronto, Canada: Author. Retrieved from http://www.edu.gov.on.ca/extra/eng/ppm/ppm8.pdf

Pearson. (2014, December). *Differences in Canadian and U.S. norms using the WAIS-IV* [Special note]. Retrieved from http://www.pearsonassess.ca/content/dam/ani/clinicalassessments/ca/programs/pdfs/WAIS-IV_Special_Note_Dec2014.pdf

Pfeiffer, S. (2012). Current perspectives on the identification and assessment of gifted students. *Journal of Psychoeducational Assessment*, *30*(1), 3-9.

Plante, T. G. (2010). *Contemporary clinical psychology* (3rd ed.). Hoboken, NJ: John Wiley.

Ryan, J., & Schnakenberg-Ott, S. (2003). Scoring reliability on the Wechsler Adult Intelligence Scale–Third Edition (WAIS-III). *Assessment*, *10*, 151-159. doi:10.1177/1073191103252348

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children*. New York, NY: The Psychological Corporation.

Wechsler, D. (1991). *The Wechsler Intelligence Scale for Children–Third Edition*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1996). *WISC-III manual Canadian supplement*. Toronto, Ontario: Harcourt Brace Canada.

Wechsler, D. (2001). *Wechsler Adult Intelligence Scale–Third Edition: Canadian technical manual*. Toronto, Ontario: Harcourt Canada.

Wechsler, D. (2004a). *The Wechsler Intelligence Scale for Children–Fourth Edition*. London, England: Pearson Assessment.

Wechsler, D. (2004b). *Wechsler Intelligence Scale for Children–Fourth Edition: Canadian technical manual*. Toronto, Ontario, Canada: The Psychological Corporation.

Wechsler, D. (2008a). *Wechsler Adult Intelligence Scale–Fourth Edition: Canadian Technical Manual*. Toronto, Ontario: Pearson Canada.

Wechsler, D. (2008b). *Wechsler Adult Intelligence Scale–Fourth Edition: Technical and Interpretive Manual*. San Antonio, TX: Pearson.

Wechsler, D., & Naglieri, J. A. (2006). *Wechsler Nonverbal Scale of Ability*. San Antonio, TX: Harcourt Assessment. doi:10.1177/0734282908329108

Weiss, L., Keith, T., Zhu, J., & Chen, H. (2013). WAIS-IV and clinical validation of the four- and five-factor interpretative approaches. *Journal of Psychoeducational Assessment*, *31*(2), 94-113.

World Health Organization. (2010). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*. Geneva, Switzerland: Author.

Zachary, R. A., & Gorsuch, R. L. (1985). Continuous norming: Implications for the WAIS-R. *Journal of Clinical Psychology*, *41*, 86-94.

Zhu, J., & Chen, H. (2011). Utility of inferential norming with smaller sample sizes. *Journal of Psychoeducational Assessment*, *29*, 570-580. doi:10.1177/0734282910396323