Journal of Psychoeducational Assessment
2015, Vol. 33(4) 312–328

DOI: 10.1177/0734282915578577
jpa.sagepub.com
SAGE

Intelligent Use of Intelligence Tests: Empirical and Clinical Support for Canadian WAIS–IV Norms

Jessie L. Miller[1], Lawrence G. Weiss[2], A. Lynne Beal[3], Donald H. Saklofske[4], Jianjun Zhu[2], and James A. Holdnack[2]

Abstract

It is well established that Canadians produce higher raw scores than their U.S. counterparts on intellectual assessments. As a result of these differences in ability along with smaller variability in the population's intellectual performance, Canadian normative data will yield lower standard scores for most raw score points compared to U.S. norms. Two recent studies have questioned the utility of the WAIS–IV Canadian norms based on the performance of a mixed clinical sample of post-secondary students. These studies suggest that a greater proportion of cases from their mixed clinical samples fall below a full-scale IQ of 85 using the WAIS–IV Canadian norms than should be "expected." The purpose of the current study is threefold: First, to summarize the consistent finding of Canada–U.S. differences on measures of ability and present new empirical analyses to demonstrate these results are not due to a smaller sample size for Canadian norms. Second, and most importantly, matched sample comparisons demonstrate that the proportion of low scoring individuals (FSIQ < 85) in mixed clinical samples is consistent with the rates published by recent studies, and not greater than expected. Third, we offer evidence-based advice to clinicians practicing in Canada on the appropriate use of Canadian norms for Canadian clients during an individual assessment of intellectual functioning.

Keywords

WAIS–IV, Canadian norms, intelligence, diagnosis, learning disability, intellectual disability, standards, clinical assessment, Canadian adults, advice to practitioners, IQ

Canadian norms for the Wechsler intelligence tests have been developed for psychology professionals working with clients from the Canadian population since the Wechsler Intelligence Scale for Children third edition (WISC–III) Canadian standardization study (Wechsler, 1996). The development of specific norms for Canada for American-based intelligence tests was a response

[1]Pearson Clinical Assessment, Toronto, Ontario, Canada
[2]Pearson Clinical Assessment, San Antonio, TX, USA
[3]Private Practice, Toronto, Ontario, Canada
[4]University of Western Ontario, London, Canada

Corresponding Author:
Jessie L. Miller, Pearson Clinical Assessment, 55 Horner Avenue, Toronto, Ontario M8Z 4X6, Canada.
Email: jessie.miller@pearson.com

to urging from Canadian practitioners who argued that U.S. norms were not representative of Canadian performance (Beal, 1988). An earlier study by Holmes (1981) showed higher mean scores and smaller variability on the verbal intelligence quotient (VIQ; $M = 104.33$, standard deviation [SD] = 12.98), the performance intelligence quotient (PIQ; $M = 108.83$, $SD = 13.12$), and the full-scale intelligence quotient (FSIQ; $M = 106.95$, $SD = 12.66$) of the WISC–R across a sample of 340 Canadian children ages 7, 9, and 11 years living in British Columbia (B.C.). The use of U.S. normative data yielded higher than expected ability scores and supported practitioner requests for Canadian normative data (Holmes, 1981). Based on these findings, Holmes made an adjustment to the WISC–R U.S. norms and created conversion tables for IQ by anchoring the mean at 100 and the *SD* at 15. As a result, these adjusted B.C. norms yielded lower scaled scores compared with using U.S. norms, and this difference was more pronounced in the lower score ranges. Indeed, more than 5 times as many children scored 70 or lower on FSIQ using B.C. norms than using the U.S. norms in Holmes's sample. Still, the proportion of cases <70 using B.C. norms corresponded to the proportion expected based on the normal curve distribution, whereas the rate of cases <70 using U.S. norms grossly underestimated the expected proportions.

Clinical anecdotes of differences in performance among Canadian examinees scored using U.S. norms continued to permeate discussions among Canadian practitioners. In the interim between the publication by Holmes in 1981 and the eventual publication of the first set of Wechsler Canadian norms, many practitioners were faced with discordant opinions on how to use WISC–R norms when assessing Canadian children. Some applied an adjustment based on the B.C. data reported by Holmes (e.g., Hardman & Oldridge, 1985), some advocated continued use of the U.S. norms (Dash et al., 1983) and still others argued for Canadian norms (Beal, 1988). A consistent approach to describing the cognitive performance of Canadian children was sorely needed.

When the first representative Canadian norming study occurred in 1996 on the WISC–III, results confirmed earlier findings reported in Holmes (1981). Performance levels were higher and the distribution was smaller among Canadian children compared with American children in the WISC–III normative samples. The overall difference in performance between the Canadian and U.S. samples on the WISC–III was 3.34 full-scale points, 1.94 verbal IQ points and 4.96 performance IQ points, when U.S. norms were applied (Wechsler, 1996). As described in the WISC–III Canadian Manual (Wechsler, 1996), using Canadian norms resulted in lower FSIQ, VIQ, and PIQ scores compared with U.S. norms. This difference was noted as being greater in the high and low ranges of the test where differences in the population distribution were the greatest (Wechsler, 1996).

Based on the results of the WISC–III Canadian standardization study, performance differences were investigated several years later on the Wechsler Adult Intelligence Scale third edition (WAIS–III). On the WAIS–III, Canadians obtained higher raw scores and exhibited smaller performance variability than American adults. Differences between the standardization samples in the United States and Canada are reported in the WAIS–III Canadian Technical Manual (Wechsler, 2001). While demonstrating that the WAIS–III was assessing the same cognitive constructs in both the United States and Canada, an analysis of the raw scores and distributions supported the need to develop independent Canadian WAIS–III norms (Saklofske, Patterson, Gorsuch, & Tulsky, 2001).

Since the publication of the WISC–III and WAIS–III Canadian standardization studies, Wechsler Canadian norms continue to be developed and used for Canadians in the assessment of intellectual ability as well as academic achievement (Pearson, 2010; Wechsler, 2001, 2004a, 2004b, 2008a, 2012, 2014a). The need for Canadian norms is substantiated by consistent findings of Canadian–U.S. performance differences on Wechsler ability and achievement tests, albeit with some variability in the oldest adult age groups (Saklofske et al., 2001). For the last 20 years of Wechsler test development in Canada, regardless of age, sample cohort, sample size, the norming

methodology or the structure of the test, Canadians continue to demonstrate higher average raw score performance than Americans. This robust finding makes it difficult to attribute these differences simply to sampling error or inadequate statistical power.

The recommendation to use WAIS–IV Canadian norms when testing Canadian adults has been challenged by Harrison, Armstrong, Harrison, Lange, and Iverson (2014) and Harrison, Holmes, Silvestri, and Armstrong (2015). The authors report that Canadian norms produce lower standard scores compared with American norms in mixed clinical samples of Canadian adults. The central hypothesis and subsequent results of both articles by Harrison et al. rest on the finding of statistically significant differences across all IQ, index, and subtest scaled scores when comparing Canadian and U.S. norms on the same individual and furthermore, that the Canadian norms consistently produce lower scores. These findings are unremarkable on their own given the previously established and published data on this difference. Every Canadian Wechsler ability test adapted since WISC–III has published these differences in their respective Canadian manuals. Harrison et al. make further claims as to the explanation behind these differences (e.g., small samples) and that the degree of difference is greater than is expected for this particular population (i.e., more low scoring cases using Canadian norms than expected) that warrant examination and addressing. Furthermore, the tenet made in these articles that "clinically meaningful" differences occur based on the choice of normative set applied reflects a rigid adherence to cut-scores on IQ tests that deserves renewed discussion in the field for all concerned with best practices in clinical assessment in Canada, and in general.

In the current study, we reconsider data reported by Harrison et al. (2014, Harrison et al., 2015) in the context of expected score distributions for mixed clinical samples of young adults, taking into account the impact of a long-standing pattern of higher raw score means in numerous Canadian normative samples, and arrive at different conclusions about the utility of the WAIS–IV Canadian norms.

The Current Study

This study has three components; the first component describes why the Canadian norms produce lower scaled scores than the American norms on the WAIS–IV. Here, examples are presented to illustrate the impact of a smaller *SD* and higher mean on Canadian scaled scores versus U.S. scaled scores. In addition, new results using the WAIS–IV data show the non-significant impact of sample size on these country differences. The second component tests the claim from the Harrison et al. (2014, Harrison et al., 2015) studies that the proportion of their clinical samples scoring below specific FSIQ cut-points using Canadian norms is higher than that should be expected, and related to this, that the percentage of cases scoring at these same cut-points using the U.S norms is a more accurate depiction of their sample. Data from the WAIS–IV standardization studies show that the results obtained by Harrison et al. are in fact, consistent with expectations based on the distribution of intelligence scores in a mixed clinical sample, and contrary to claims by Harrison et al., offer additional evidence of the validity of the Canadian norms. The third component addresses the practical issue of test interpretation, classification, and diagnosis using cut-scores, and provides best practice recommendations for using Canadian norms.

Differences in WAIS–IV Canadian and U.S. Norms

The raw score distributions of the WAIS–IV Canadian normative sample are different from the WAIS–IV U.S. normative sample. Across all age groups and most subtests, Canadians obtained higher raw scores than Americans in the normative sample. At the overall level, Canadians obtained a mean FSIQ of 104.51 points compared with the U.S. mean of 100, when both samples were scored using U.S. norms. In addition, the *SD* was smaller for Canadians than the U.S. by 1.6

Table 1. Sample Raw Score to Scaled Score Comparison Using Canadian Versus American WAIS–IV Norms.

Example 1: Difference in *SD* only	Example 2: Difference in *M* and *SD*
Canadian population *M* of 20	Canadian population *M* of 21
U.S. population *M* of 20	U.S. population *M* of 20
Individual raw score of 14	Individual raw score of 14
Canadian *SD* = 2, U.S. *SD* = 3	Canadian *SD* = 2, U.S. *SD* = 3
Canadian *z score* = (14 − 20) / 2; then *z* = −3, thus	Canadian *z score* = (14 − 21) / 2; then *z* = −3.5, thus
Canadian standard score = (−3) × 15 + 100 = 55.	Canadian standard score = (−3.5) × 15 + 100 = 47.5.
In contrast, U.S. *z score* = (14 − 20) / 3 = −2, thus	In contrast, U.S. z score = (14− − 20) / 3 = −2, thus
U.S. standard score = (−2) × 15 + 100 = 70.	U.S. standard score = (−2) × 15 +100 = 70.

Note. WAIS–IV = Wechsler Adult Intelligence Scale–IV.

points when both normative samples were scored using U.S. norms. These performance characteristics combine to produce greater differences on the resulting raw to scaled score conversions than would be expected for the mean difference in isolation.

Based on the mean difference, given the same low raw score, the converted standard score is further away from the mean in the Canadian sample than in the U.S. sample. Furthermore, these Canada–U.S. WAIS–IV differences do not generalize in a simple linear fashion across age, subtest or ability level. Variability is smaller below the mean than above the mean for the Canadian sample and this occurs both at the scaled score level and raw score level. At the raw score level, the Canadian scores are more negatively skewed than U.S. scores across the core FSIQ subtests and this is both in the number of skewed subtests, and in the size of the negative skew. For this reason, performance discrepancies at the full-scale level tend to be greatest in the lower tail of the distribution rather than occurring to the same degree throughout the distribution.

With a smaller *SD* and a larger negative skew, the distribution of intelligence scores in the WAIS–IV Canadian sample becomes narrower and scores drop off more quickly at the lower tail relative to the U.S. curve. As a result, the difference between a typical and an atypical ability case is greater using the Canadian norms than the U.S. norms. The following hypothetical examples illustrate the impact that a smaller *SD* has on the raw score to scaled score conversion. Example 1 limits the impact to a smaller *SD* in the Canadian norms. Example 2 shows that, with the combined differences in population's means and *SD*s working together, the standard score differences can be even greater (Table 1).

These hypothetical examples demonstrate the extreme score variation that can occur with a one-point difference in *SD* and a one-point difference in the mean score between samples which is smaller than the actual difference between the two country means. These differences in the distribution of raw scores are the primary reasons for the observed differences in WAIS–IV scaled scores between the two countries.

It is important to note that beyond a comparison at the FSIQ level, meaningful subtest and primary index differences exist between the WAIS–IV Canadian and American norms. The mean scaled and standard scores for subtests and composites are not 10 or 100 (the expected values) when the U.S. norms are used (see Table 4.6 of the WAIS–IV Canadian Manual, Wechsler, 2008a). When using the U.S. norms, the range of Canadian mean scaled scores range from 10.1 to 11.0 for subtests, and 102.3 to 105.0 for composites. These differences affect inter-subtest and index scatter so that the resulting profile analysis may be difficult to interpret when U.S. norms are used because such an analysis assumes all standardized scores are scaled to the same mean.

While the exact causes of these population differences are unknown, mean differences in performance between Canada and the United States on the WAIS–IV, as well as on other Wechsler Intelligence tests, are often attributed to important variations in the demographic composition of the two countries. The largest demographic contributor to performance on measures of ability is education level (Kaufman & Doppelt, 1976; McDermott, 1995; Sattler, 2001; Wechsler, 2012). The census-based sample targets between Canada and the United States on WAIS–IV are notably different in education level. Specifically, the total proportion of the WAIS–IV normative sample attending college for 1 or more years is 56.7% in Canada compared with 43.9% in the United States. The United States and Canada also differ in ethnic diversity with the United States being more diverse than Canada. These demographic differences indicate a more homogeneous and more highly educated population from which to draw a normative sample in Canada relative to the United States. This may account for the smaller variability and higher mean scores observed in the WAIS–IV Canadian data.

The matched sample analyses reported in the WAIS–IV Canadian manual tends to support demographic differences being a primary explanation for ability score differences between Canada and the United States. When WAIS–IV Canadian and U.S samples are matched on age, sex, ethnicity, and parent or self-education level, the mean score differences decreased or were eliminated completely. Where significant differences continued to be present after matching, effect sizes were small, ranging from .11 to .31. When the sample was divided by age group, 16- to 69-year-olds showed small mean differences compared with the U.S. matched sample, whereas differences disappeared completely in the oldest age group (70- to 90-year-olds). Comparative analyses between Canadian and U.S. normative samples on other Wechsler scales including prior editions of WISC and also the Wechsler Preschool and Primary Scale of Intelligence (WPPSI) have yielded consistent results: When the normative samples are matched on demographic characteristics, the Canada–U.S. score differences are significantly reduced or disappear altogether.

This finding with matched Canadian and American samples has been replicated with the Woodcock–Johnson III NU (WJ-III; Ford, Swart, Negreiros, Lacroix, & McGrew, 2010). Ford et al. (2010) found that by equating the two countries on demographic characteristics, the mean differences in cognitive ability test scores between the two samples became non-significant. Furthermore, their data also showed that the *SD* of the Canadian sample (13.52) was much smaller than the *SD* of the matched American sample (15.72) The Ford et al. results demonstrate that it is the functioning of the individual within the context of his or her country's demographic characteristics that makes for the differences in the scores obtained.

These matched sample comparison studies provide insight into factors that may relate to score differences between the two populations. This is not to imply that the explanation of the performance differences between Canada and the United States is singular or easily explained. However, they provide plausible reasons that are theoretically and empirically supported. Furthermore, the established measurement equivalence of the WAIS–IV across U.S. and Canadian standardization samples confidently rules out the possibility that mean differences are attributable to measurement error (see Bowden, Saklofske, & Weiss, 2010).

With similar results in their studies using a referred sample, Harrison et al. (2014, Harrison et al., 2015) have suggested that small Canadian sample sizes may explain the occurrence of lower scaled scores in Canada compared with that in the United States. This suggestion has not been supported in empirical studies and is not consistent with the robust finding of U.S.–Canadian differences across varying sample sizes (see Wechsler, 1996, 2001, 2004a, 2008a, 2012, 2014a). Canadian sample sizes are typically 75 to 100 per age band across normative studies, although in certain age bands they have been as low as 30. These sample sizes are approximately half to one quarter the size of the sample per age bands used in the United States. Canadian samples are smaller predominantly because Canada has a more homogeneous population with fewer demographic variables to represent and thus does not require the same number of participants to

generate a sampling stratification matrix which accounts for all potential interactions among demographic variables within each age band. Nevertheless, while the Canadian samples are notably smaller than American samples, lower scaled scores are consistently found using Canadian norms and this occurs regardless of whether there are 30 or 100 cases per age band.

Further Evidence of the Validity of the Canadian WAIS–IV Norms

Our current study has two goals. The first goal is to test the claims made by Harrison et al. (2014, Harrison et al., 2015) that the number of individuals in their samples who score below a particular cut-score on FSIQ using the Canadian norms was far greater than should be expected. In particular, in both articles they suggest that a far greater number of individuals fall below an FSIQ of 85 using Canadian norms than using American norms, and that these differences lead to clinically meaningful outcomes for the individuals being assessed. We test these claims by examining the distribution of FSIQ scores on the WAIS–IV among two matched comparison samples that parallel the clinically referred sample of adults seeking post-secondary program accommodation reported by Harrison et al. (2014, Harrison et al., 2015). The second goal is to empirically investigate the impact of reduced sample sizes on the accuracy of norms.

Method

All results presented are based on analyses using the data from the WAIS–IV U.S. standardization studies (Wechsler, 2008b). We derived two mixed clinical samples using the U.S. WAIS–IV standardization and validation studies which include both clinical and normative cases (a full description of the clinical and nonclinical samples including data collection procedures and inclusion/exclusion criteria can be found in the WAIS–IV Technical and Interpretive Manual, Wechsler, 2008b).

We selected our sample to match the sample characteristics specified in the Harrison et al. (2014, Harrison et al., 2015) studies in two ways: We first restricted the cases available for random selection to the age range described in the 2014 and 2015 studies. Next, while years of education and socio-economic background of the participants in the Harrison et al. samples were not reported, we matched by education level as closely as possible by including predominantly education Level 3 cases from the WAIS–IV sample. Education Level 3 corresponds to 12 years of schooling or a high school diploma. More than 90% of cases in our matched samples had 12 years of education or greater. Cases where education was ≤8 years or ≥16 years (college or graduate degree) were not eligible for inclusion. A small percentage of cases with education Level 2 (9-11 years) were included to correspond to the cases in the Harrison et al. samples that were 16 to 17 years old (~3%) as well as the unreported number of mature students acknowledged in the Harrison et al. sample who, by definition in Ontario, have not completed high school. For both matched samples, cases 16 to 19 years of age were based on parent education level, per the standardization procedures of the WAIS–IV Canadian and U.S. studies (Wechsler, 2008a, 2008b). All other cases for ages 20 years and older were based on self-education level.

For our first matched sample, a proportion of nonclinical cases from the U.S. normative sample were matched on age and education as described above, to the Harrison et al. (2014) sample and $n = 60$ cases were then randomly drawn to represent 23.3% of nonclinical cases, matching the percent of undiagnosed participants in the Harrison et al. (2014) study. For the second matched sample, again matching on age and education, $n = 24$ cases were randomly drawn from the U.S. normative sample to represent 11% of nonclinical cases, matching the percent of undiagnosed participants in the Harrison et al. (2015) study. To match the clinical cases in the Harrison et al. studies, $n = 197$

and $n = 197$ were drawn from the available U.S. clinical data to represent 76.7% of cases and 89% of clinical cases per the sample characteristics published in the Harrison et al. studies.

Note that the sample used in the Harrison et al. (2015) study includes all the participants previously reported in their 2014 study plus additional participants. Thus, the two Harrison et al. studies cannot be considered to be independent replication studies. Similarly, our two matched samples contain overlapping participants, but were randomly selected to match the characteristics of each Harrison et al. sample, which differ mainly in the smaller percentage of nonclinical cases in their second study.

Results

Matched comparison studies. A total of 257 participants comprised our first matched comparison sample. The mean age of participants was 25.0 years ($SD = 13.3$ years; range = 16-57 years), and 43% were female. The ethnic background of the majority of participants was Caucasian. For the clinical cases, 28% had a diagnosis of a specific learning disability (LD), 14% had a diagnosis of attention-deficit hyperactivity disorder (ADHD), and 34.7% of the clinical cases had other diagnoses including generalized anxiety, depression, borderline intellectual functioning, and other non-LD or ADHD diagnoses. We intentionally excluded any clinical cases diagnosed with traumatic brain injury, mild cognitive impairment, mild or moderate intellectual disability (ID), Alzheimer's, or other forms of dementia, as these types of diagnoses represent more extreme impairment and we do not know for certain whether they were present in the Harrison et al. samples. Consistent with the sample characteristics reported by Harrison et al. (2014), 23.3% of our matched sample had no clinical diagnosis.

A total of 221 participants were included in the second matched sample with a mean age of 25.2 years ($SD = 12.9$ years; range = 16-63 years) and 45% were female. The ethnic background of the majority of participants was Caucasian. For the clinical cases, 32.5% had a diagnosis of a specific LD, 16.3% had a diagnosis of ADHD, and 40.2% of the clinical cases had other diagnoses including generalized anxiety, depression, borderline intellectual functioning, and other non-LD or ADHD diagnoses. No cases of traumatic brain injury, mild cognitive impairment, mild or moderate ID, Alzheimer's, or other forms of dementia were included in our matched sample. Consistent with the sample characteristics reported by Harrison et al. (2015), 11% of our matched sample had no clinical diagnosis.

The FSIQ score distributions of our U.S. matched clinical samples using U.S. WAIS–IV norms were calculated and contrasted to the FSIQ score distributions in the Harrison et al. Canadian samples using Canadian WAIS–IV norms. The descriptive statistics and percentages of FSIQ scores at various cut-points are reported for both matched samples in Tables 2 and 3 along with the respective Harrison et al. sample. As can be seen from Tables 2 and 3, the percentage of cases from our American matched comparison samples scoring below each FSIQ cut-point are very similar to the Canadian percentages reported by Harrison et al. (2014, Harrison et al., 2015).

The frequency distributions for the WAIS–IV U.S. and Canadian normative samples are also shown in Tables 2 and 3. The distributions of scores in the Canadian and American WAIS–IV normative samples are very close to theoretical expectations based on the normal curve. In contrast, the distributions of scores in the Harrison et al. studies and our own matched clinical samples are very different from normal distributions. These findings highlight the clinical composition of the Harrison et al. samples.

Sample size study. One hundred random samples were drawn from the total cases available in the U.S. WAIS–IV normative sample ($n = 2,200$). Each random drawing comprised 681 cases, to match the sample size of the Canadian WAIS–IV normative sample. For each of the 100 random samples, means and *SD*s for FSIQ were computed. Next, median scores for both the mean and

Table 2. Percentages of the Harrison, Armstrong, Harrison, Lange, and Iverson (2014) Canadian Sample and Matched U.S. Comparison Sample Scoring Below Various FSIQ Cut-Points Using Canadian or U.S. Norms on the WAIS–IV.

	Harrison et al. (2014; *N* = 432)	Matched comparison sample #1 (*N* = 253)	WAIS–IV Canadian standardization sample (*N* = 688)	WAIS–IV U.S. standardization sample (*N* = 2,200)
Norms used	Canadian	U.S.	Canadian	U.S.
FSIQ score Range				
≤70	7.7	7.5	2.8	3.0
≤75	15.1	18.2	6.9	5.9
≤80	24.4	30.4	10.7	9.8
≤85	39.4	41.5	16.6	16.9
90-109	41.8	36.4	49.1	50.3
M	91.2	89.7	99.9	100
SD	14.6	13.8	15.0	15
Range	58-133	62-126	53-139	44-141

Note. WAIS–IV = Wechsler Adult Intelligence Scale–IV.

Table 3. Percentages of the Harrison, Holmes, Silvestri, and Armstrong (2015) Canadian Sample and Matched U.S. Comparison Sample Scoring Below Various FSIQ Cut-Points Using Canadian or U.S. Norms on the WAIS–IV.

	Harrison et al. (2015; *N* = 861)	Matched comparison sample #2 (*N* = 221)	WAIS–IV Canadian standardization sample (*N* = 688)	WAIS–IV U.S. standardization sample (*N* = 2,200)
Norms used	Canadian	U.S.	Canadian	U.S.
FSIQ score Range				
≤70	9.7	8.6	2.8	3.0
≤75	21.2	19.5	6.9	5.9
≤80	32.3	32.1	10.7	9.8
≤85	48.9	44.3	16.6	16.9
90-109	34.2	33.0	49.1	50.3
M	88.1	89.0	99.9	100
SD	14.4	13.9	15.0	15
Range	Not available	62-126	53-139	44-141

Note. WAIS–IV = Wechsler Adult Intelligence Scale–IV.

the *SD* were computed. Out of 100 FSIQ means, the median was 100.1 (range = 98.6-101.4). Out of 100 FSIQ *SD*s, the median was 15.0 (range = 14.0-15.9). Out of 100 random samples, the medians for the mean and *SD* are almost identical to the theoretically expected values of $\mu = 100$ and $SD = 15$. In addition, out of 100 random samples, 86% of FSIQ means are equal to or less than 100.5 points and 97% of FSIQ means are equal to or less than 101 points. Moreover, 93% of *SD*s are equal to or greater than 14.5, and 100% of *SD*s are greater than 14.0.

These results provide empirical support that a sample size of $n = 681$ is unlikely to over-estimate the population mean by 4.51 FSIQ points (mean difference between the U.S. and Canadian

normative samples when scored using U.S. norms) or to underestimate the population *SD* by 1.6 points (mean *SD* difference between the U.S. and Canadian normative samples when scored using U.S. norms).

Discussion

This study shows that mixed clinical samples of adults can be expected to obtain low IQ scores at much higher percentages than would be expected in nonclinical samples of adults. We have replicated twice the FSIQ score distributions reported by Harrison et al. (2014, Harrison et al., 2015) using samples of *American* adults scored on U.S. norms who were carefully matched on clinical status, education level, and age, to the sample of clinically referred *Canadian* post-secondary students in the Harrison et al. (2014, Harrison et al., 2015) studies. These results indicate that the high percentage of low scoring participants in the studies by Harrison et al. appears to be the result of their samples comprising more than two thirds of individuals with clinical diagnoses.

We have shown that the proportion of cases from a referred clinical population with "below average" (FSIQ < 85) performance is consistent with the rates of 39% and 48.9% reported by Harrison et al. (2014, Harrison et al., 2015). Our first matched comparison sample scored using U.S. norms showed an equivalent percentage of low performing individuals (40.1%) as that reported by Harrison et al. (2014; 39.4%) using Canadian norms. Our second matched comparison sample scored using U.S. norms again showed a similar proportion of individuals scoring below average (44.3%) as that reported by Harrison et al. (2015; 48.9%) using Canadian norms. Most notably, neither our American samples nor Harrison et al.'s Canadian samples came close to approximating the 16% of scores which fall below an FSIQ of 85 in the Canadian and American normative samples, and which Harrison et al. (2014, Harrison et al., 2015) state to be the expected rate of below average performance. As evidenced by the data presented in Tables 2 and 3 of our results section, the Harrison et al. clinically referred samples are not comparable with normal populations. Harrison et al's samples do not follow the expected distribution of intelligence based on the theoretical curve–nor should they. Samples that primarily comprise clinically diagnosed individuals who have disorders associated with performance deficits on measures of ability are not representative population samples, and should not be expected to yield scores that approximate a normal distribution.

The precise distribution of intelligence scores expected from a mixed clinical sample is difficult to estimate given that severity of clinical status can vary considerably across individuals with the same diagnosis. However, what is clear is that an assumption that only 16% of a clinical sample should score less than 85 is inconsistent with the available evidence. Rather, the rate seems to be more accurately described as somewhere between 40% and 50% based on the clinical characteristics of the samples reported both here and in the articles by Harrison et al. Such expectations are consistent with the mean FSIQ's of our matched clinical samples and Harrison et al.'s samples which hover around 89 full-scale points. Furthermore, the current data are consistent with the mean FSIQ scores reported for LD and ADHD samples in the U.S. technical manuals of the WAIS–III, WAIS–IV, WISC–III, WISC–IV, and WISC–V (Wechsler, 1991, 1997, 2003, 2008b, 2014b). Clearly, many scores below 85 will be observed in any clinical sample in which the expected mean is around 89 or 90 points.

Our results further demonstrate that the mean and distribution of ability scores in the U.S. WAIS–IV normative sample will not change shape simply by reducing the sample size from 2,200 to a size that is identical to the Canadian WAIS–IV normative sample of 681. While a small sample size is an easy target for criticism when data are in question, the fact is the power of the Canadian sample size is adequate for modeling the distribution of ability scores. According to research published by Zhu and Chen (2011) and Wilkins, Rolfhus, Weiss, and Zhu (2005),

inferential norms developed with sample sizes of 50 per age group are as good as, or better, than traditional norms derived from sample sizes of 100 per age group.

It is important to note that the 11% and 23% of nonclinical cases included in our matched comparison samples are not equivalent to the undiagnosed cases in the Harrison et al. samples. Simply put, their undiagnosed cases are likely lower functioning than our nonclinical cases. The nonclinical cases in the Harrison et al. studies are still part of a referred population that were experiencing problems related to their academic functioning and while they did not receive a formal clinical diagnosis, they are a sub-clinical group referred to investigate self-reported LDs that are likely affecting their performance on a wide variety of academic and cognitive tasks. The nonclinical cases in our matched samples had not been referred for evaluation due to academic, cognitive, or other problems, and in fact, these conditions were exclusionary criteria for participation in the normative sample.

Several other differences between the samples in the Harrison et al. studies and our matched comparison samples should be noted as they may collectively account for significant differences in overall performance. For instance, it is not known whether the undiagnosed cases in Harrison et al. samples were taking any psychotropic medicated at the time of testing that would depress performance. Both the Canadian and American WAIS–IV normative samples were carefully monitored for medication use among nonclinical participants and anyone who took medication known to have significant effects on cognitive performance at the time of testing were excluded from the final sample (see pp. 40-41 of the WAIS–IV Canadian Manual and pp. 31-32 of the WAIS–IV Technical and Interpretive Manual for more details).

Another difference between our matched comparison samples and the samples in the Harrison et al. studies is that all WAIS–IV participants had to meet strict English language inclusion criteria to be part of the normative or clinical samples. All participants in the WAIS–IV standardization had to either have English as their primary language or demonstrate proficiency in English if they were multi-lingual. The English language proficiency of the participants in the studies by Harrison et al. is unknown. It is worth highlighting that Ontario attracts more than 40% of Canada's large and growing population of international post-secondary students each year, most of whom come from countries where the primary language is not English (Prairie Research Associates Inc, 2009). Indeed, the bulk of these international students arrive in Ontario for enrollment in language training courses offered through Ontario's colleges (Kunin & Associates, Inc, 2012).

Another very important distinction between our comparison samples and the Harrison et al. referred samples was the goal of the assessments. The normative and clinical cases collected during standardization are collected for research purposes, and while participants are paid a small monetary incentive for participating, there was no external motivation to over or under perform on the WAIS–IV. In sharp contrast, the participants in the studies reported by Harrison et al. were seeking valuable educational accommodation from their respective post-secondary programs, including possible disability tax credits or income benefits from the government, which may have incented some students not to give their best effort. Harrison et al. (2015) acknowledged this likelihood when they report administering measures of effort, or performance validity, to all participants, yet they did not report the results of the effort measures in either the 2014 or 2015 articles. They argue that any possible reduced effort is not material because each participant serves as their own control when their test performance is scored on Canadian and American norms and then compared. However, reduced effort can clearly contribute to an overall lowering of the raw score performance. As described above, the smaller Canadian *SD* results in any low raw score being further below the Canadian than American mean in standard score units. Thus, any possible reduced effort not only lowers the raw and standard scores but increases the difference between the Canadian and American derived scores. Given the large degree of secondary gain that would result from a reduced level of effort on this test, any conclusions drawn from the

performance of participants in the Harrison et al. samples are questionable without evidence that participants put forth their best effort on the test. However, evidence of performance validity, even if subsequently presented, would not obviate the numerous other methodological criticisms noted here.

The socio-economic status (SES) of the Harrison et al. samples is also unknown. SES is a powerful predictor of intelligence test scores, but no information is provided about the percent of students who received government assistance to help pay for the evaluation. SES is often approximated using educational attainment for adults, and yet the educational backgrounds of the Harrison et al. samples are also unknown. First, 13 participants in the 2014 Harrison et al. study sample and 32 participants in the 2015 Harrison et al. study sample were ages 16 to 17 years. This age is not consistent with the age of entry into post-secondary programs in Ontario (18 years) and suggests the sample comprised a small percentage of current high school students. The inclusion of these participants in samples that are deemed to be 'currently enrolled in post-secondary programs' is unusual and unclear. Second, in describing the participants in their sample, Harrison et al. (2014) stated that "a sizeable minority were returning as mature students." In Ontario, mature students are defined as those above the age of 19 years who have not completed high school or a GED. This means that an unknown proportion of the Harrison et al. samples had not graduated from high school but yet were attempting to transition to post-secondary programs which they may not have been prepared for. Given the range of programs that are offered through Ontario colleges and universities, the lack of detail provided on program types or admissions criteria is also curious, particularly given the relevance of this information in establishing functional levels of the samples in the Harrison et al. studies. Ontario colleges and universities offer a wide range of certificates, diplomas, and degrees, and they vary in length from mere weeks to years of training. The various technical versus academic skills required for admission also vary widely by program type. Harrison et al.'s description of their samples as post-secondary students who "graduated from high school with marks high enough to qualify for acceptance into bona fide post-secondary programs" is not verifiable from the available data. The high school grade point average (GPA) of the students is unknown, the number of students with completed high school degrees is unknown, and the type of program and admission requirements needed for their current enrollment status is also unknown, not to mention the fact that some of the participants were referred specifically based on "weak academic performance."

In summary, the data from the Harrison et al. (2014, Harrison et al., 2015) studies are consistent with previous comparisons of Canadian and American norms for WISC-R, WISC-III, WISC-IV, WISC-V, WAIS-III, WAIS–IV, WPPSI-III, and WPPSI-IV: Canadian norms yield lower scaled and standard scores then U.S. norms when compared across the same individuals. The Canadian norms are different than the U.S. norms because the raw score performance of Canadians is different. It was this finding in the Holmes (1981) study in B.C. and in the Canadian validation study of WISC-III (1996) that led to the development of Canadian norms. This is what Canadian practitioners wanted (Beal, 1988), and what the data indicated was the right thing to do.

Reinterpreting the data in light of expected scores for mixed clinical samples of young adults, and the effect of higher raw scores combined with smaller population variances on Canadian WAIS–IV standard scores, we arrive at the opposite conclusion from Harrison et al. (2014, Harrison et al., 2015): We believe that the Canadian IQ scores reported in the Harrison et al. articles are consistent with the expected distribution of IQ scores in mixed clinical samples, and we have presented supporting evidence from matched clinical samples. We further believe that the sometimes large differences between scores derived from American and Canadian norms on the WAIS–IV are consistent with differences between the populations in terms of mean ability and homogeneity of variability, and we have presented evidence accordingly. Thus, we conclude that the Harrison et al. data actually provide evidence that Canadians who take the WAIS–IV should be scored with the Canadian norms.

General Discussion

The best representation of an individual's intellectual ability is from a psychometrically sound test that uses a population norm that is drawn from the *same* population as the individual being assessed. These are the guiding principles and standards of assessment. The practice of using U.S. norms to evaluate the intellectual ability of Canadians is inconsistent with the Principles for Fair Student Assessment Practices for Education in Canada (Principles; Joint Advisory Committee of the Canadian Education Association, 1993) and the Standards for Educational and Psychological Testing (Standards; American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014) which are also endorsed by the Canadian Psychological Association. Both the Principles and the Standards stipulate that best practice guidelines in student assessment must include normative data that are derived from and representative of the same population as the student being assessed. Comparing relative placement of Canadians on the U.S. norms versus the Canadian norms without sufficient justification for doing so is counter to best practice guidelines. Standard IQ scores must be interpreted vis-à-vis one's reference group and comparing scores across reference groups assumes the populations are the same. However, we know from a number of studies that the Canadian population performs differently on ability measures and we know the distributional properties are different across Canada and the United States. These are known, documented, published, statistical facts.

Clinicians must be cautious when interpreting test scores against criterion measures that were established in absentia of their own population. Indeed, the unique distributional properties and ability level of Canadians have never been incorporated into American standards for diagnosing ID. The statistical characteristics of the Canadian population on measures of IQ were largely unknown until the mid-1990s when the first large and representative sample of Canadian performance on the WISC–III was published by the Psychological Corporation (Wechsler, 1996). In fact, the origin of the IQ cut-score above and below which clinicians rule out a diagnosis of ID came out of the need to establish benchmarks for practitioners working in the United States. In 1961, the American Association on Intellectual and Developmental Disabilities (AAIDD, formerly the AAMD) set a ceiling for ID based on the perception that the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM–II*; APA, 1968) lacked scientific merit in the ID criteria (Greenspan & Woods, 2014). Based on a normal distribution of intelligence with a mean of 100 and *SD* of 15 points, the AAIDD established the ceiling for identifying ID at 1 *SD* below the population mean (IQ of 85). This ceiling was set purposely high with the intention that it was better to be over-inclusive in the identification criteria because the adaptive behavior criteria could be used to narrow the pool of those most in need of services. However, a lack of adherence to the adaptive behavior criterion and an over-reliance on the IQ criterion led to an over-identification of individuals as ID and thus the AAIDD further reduced this ceiling to an IQ of 70 in their 1973 publication (Greenspan & Woods, 2014). This ceiling of 70 for a diagnosis of ID was adopted by the American Psychiatric Association (APA) in 1980 and has since become synonymous with a diagnosis of ID. While this cut-score was introduced to provide consistency among diagnoses and a framework for determining service provision among practitioners in the United States, it became an official criterion that is rigidly and exclusively followed without consideration to the parameters that many scientists and clinicians have espoused for decades; IQ scores must be interpreted within the range of measurement error of the test used and it must be interpreted within a functional assessment of behavior (see Saklofske, Reynolds, & Schwean, 2013).

The association between ID and an IQ score <70 was established on the assumption of a normal distribution with a mean of 100 and a *SD* of 15. Neither of these statistical properties reflects the population characteristics of the WAIS–IV Canadian normative sample when scored using U.S. norms. Furthermore, the confidence intervals around scores on IQ tests assume a normal distribution with a *SD* of 15. Choosing to use U.S. norms to interpret the test scores of a Canadian

will increase the variance and subsequently increases the CIs around the true score making it more difficult to rely on the test scores for diagnosis. The error associated with measuring IQ and the over-reliance on IQ scores for classification is one of the reasons the APA has modified the criteria associated with ID diagnosis in the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*; APA, 2013; Greenspan & Woods, 2014). Severity must be determined based on adaptive functioning and, while the diagnosis must include IQ measures, the scores should be interpreted alongside measures of adaptive behavior (APA, 2013). More important, the cutoff criterion for ID is "approximately two *SD*s below the mean, *but in consideration of the standard error of measurement for the specific assessment instruments used and the instruments' strengths and limitations*" (emphasis added; APA, 2013; Schalock et al., 2010).

The articles by Harrison et al. (2014, Harrison et al., 2015) left Canadian practitioners with a dilemma: With Canadian norms producing lower scores than American norms, they conclude that WAIS–IV Canadian norms over-diagnose disability among Canadians. While their data replicate other studies, their conclusion does not follow. WAIS–IV Canadian norms do result in lower standard scores than American norms—this fact has never been up for debate. However, the outcome of this difference will never be over-diagnosis of Canadian adults unless clinicians are relying exclusively on conventional IQ cut-scores for diagnosis, to the exclusion of all other clinical information, and with disregard to the known distributional and mean differences of the Canadian population's intellectual ability. If this is indeed what Canadian practitioners are doing, as argued by Harrison et al. (2015), the practice of clinical assessment in Canada does indeed have a problem, but that problem is not the accuracy of the WAIS–IV Canadian norms.

We conclude our article with practical advice and recommendations for Canadian practitioners assessing Canadian clients using WAIS–IV Canadian norms.

Best Practice Recommendations: Clinical Implications and Applications

The assessment of adults referred with a suspected learning or ID requires care in interpreting current test results. The assessment process is an integration of information across multiple sources and not an arbitrary assignment based on a single test score. Clinical assessment is the outcome of expert hypothesis testing during which the clinician attempts to determine the best interpretation of the developmental and functional trajectory of the person, up to the current time. Any diagnosis would follow from this integration, along with recommendations to accommodate learning, if needed. The clinician goes far beyond the job of a technician who would simply give a test, such as WAIS–IV and report a score.

When using WAIS–IV for assessments in Canada to determine whether an adult has a LD or an ID, the following practices are advisable.

1. Look closely at the adult's learning history. Relevant history includes developmental history, educational history, and previous assessments, including psychological, occupational therapy, speech-language, and psychiatric evaluations. Learning disabilities and IDs are lifelong conditions which are typically evident in childhood, although they might not have been detected or formally diagnosed. Ask if there is a history of early and long-standing deficits in academic skills, below the level predicted by the person's cognitive ability, linked to information processing deficits? If there is evidence of or an early history of LD or ID, then hypothesize an LD or ID continuing into adulthood. Consider also that, among low functioning individuals, achievement is typically predicted to be somewhat above IQ due to regression to the mean. See Table E.1 in the WAIS–IV Canadian Manual for examples (using an FSIQ of 80, for example).

2. Plan and conduct an assessment battery that will cover all of the elements of academic skills and cognitive abilities that could be related to an LD or ID.
3. When testing adults in Canada, use Canadian norms for tests, where available. Canadian norms reflect how the adult compares with other Canadian adults going to school or working within a Canadian context. They are the most appropriate norms for making diagnostic decisions about Canadian adults. Using U.S. norms could result in erroneous conclusions at the subtest and index levels. As noted earlier, the inter-subtest and index scatter assumes no differences between the means of scaled and standard scores. However, when Canadians are scored using U.S. norms, their subtest and composite means are no longer 10 and 100. The interpretation of the statistical difference between discrepant subtests and index scores will no longer be valid.
4. Is secondary gain a relevant consideration in this circumstance? Consider whether the adult sustained an appropriate level of effort throughout the assessment. Include at least two symptom validity tests to obtain objective measures of the adult's effort during the assessment. If the scores are above cutoffs for adults, then the cognitive and academic ability scores may be interpreted as valid and reliable indications of current abilities. If the effort measures are below cutoffs for adults, consider carefully what the impact could be on the cognitive and academic ability scores. It is possible that the test scores are of little if any clinical utility if they cannot be taken as valid and reliable indicators of the student's ability. Yet, it is important to put these results into the context of early history and current functioning. If the adult clearly met diagnostic criteria for an LD in childhood or adolescence, then it is unlikely that the LD has resolved by adulthood. Still, it is likely that the scores obtained by the adult in this assessment with unreliable effort will be below the levels of his or her actual functioning. Check the person's functional level in their past educational programs, in their current educational program, if applicable, and in their job functioning. Do the test scores reflect impossibly low results, given indications of higher levels of daily functioning? This could be particularly important for diagnostic consideration when the results are close to the boundaries of scores for different categories. Examples would include going from average to below average and from below average into an extremely low range where the consideration of a more global disability might be necessary. Consider carefully the possibility that compromised effort in the current assessment could account for current scores that are considerably below the levels of functioning seen in childhood. If the scores seem impossibly low, it may be necessary to invalidate the results and refrain from drawing conclusions on the basis of these test scores. Program modifications and supports, such as clarifications, prompts, and re-direction by teachers in high school may have historically resulted in higher marks than the student is able to obtain independently in his or her current situation.
5. Compare the adult's current scores with those obtained in childhood or adolescence. If this is a re-assessment, compare the adult's current scores with those obtained in a previous assessment and use consistent normative data sets for this comparison. Where Canadian norms are available for earlier assessments done in childhood or adolescence, continue to use Canadian norms. Use the confidence interval (CI) around the previous score and the current score in the calculation to determine if any difference is statistically significant. The severe discrepancy calculation is explained on the Dumont and Willis (2002) website (http://alpha.fdu.edu/psychology/WISCIV_DWI.htm). Apps for making this calculation are available on the Internet. If the difference is significant, then the difference between the scores reliably reflects a change in the adult's functioning over time. If it is not significant, then the difference between the scores cannot rule out error variance inherent in each test as the reason for the difference. Remember that a

large difference might not be significant, depending on the reliability of each score being compared. When the score is significantly lower now than in childhood, consider several hypotheses: Is there any academic history consistent with functioning not keeping progress with that of average persons? Is there a good reason to consider a cumulative deficit, the deficit getting larger over time? Stanovitch (1986) found that the Matthew Effect where "the rich get richer," also shows that "the poor get poorer" when it comes to reading disabilities. Children who have reading disabilities tend to show a plethora of cognitive deficits, which would undoubtedly show as lower scores on ability scales across the life span. Consider also whether there has been a neurological insult that would account for failure to continue developing at the rate observed earlier.

6. Consider carefully scores that fall in the Extremely Low range before making a diagnosis of ID. There can be several reasons for an adult scoring in the Extremely Low range on the WAIS–IV, not all of which reflect an ID. Several factors need to be considered. Does the adult have a history consistent with a developmental disability that was evident before age 18 years? If there is no previous assessment, consider anecdotal reports of how the person handled activities of daily living as a child. Consider also the educational history; was there a history of delayed progress or identification within the school district as a student with a developmental delay. Some school districts might have characterized the student or placement as "mild developmental disability (MID)."Consider the adult's current adaptive functions and activities of daily living. Inquire about their living arrangements, level of independence with home functioning, getting around in the community, and handling personal finances. If difficulties are suspected, complete an adaptive behavior scale using a close relative or partner as an informant. A diagnosis of ID would need to be supplemented by deficits in adaptive functions. Just because an adult's Full Scale IQ score or General Ability Index score falls below the cutoff, an ID could not be confirmed unless all diagnostic criteria have been met. Only with such a diagnosis would the adult be eligible for the clinician to sign off for consideration of a disability tax credit or a disability pension.Consider the adult's current activities, school program, or employment. Is the person performing in a manner that is consistent with the test scores? If performance exceeds predictions based on test scores, then test scores are likely underestimates or reflect cumulative deficits in functioning, rather than a developmental disability. The effect of learning problems accumulate also because the student does not develop at the same rate as normal peers, thus falling further behind peers with age.
7. Conclude with the most consistent explanation of the person's functioning across the life span. Integrate information across the life span with the current rest results. Test scores in adulthood reflect the trajectory of the person's cognitive and academic development from childhood, as influenced by educational opportunity and other enabling conditions. Scores may be affected by disabling conditions, such as head injury or other neurological incidents or by inadequate effort. An interpretation that considers the developmental course along with the current test results will provide the best conclusion or diagnosis.

Declaration of Conflicting Interests

The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Jessie L. Miller, Lawrence G. Weiss, and Jianjun Zhu are current employees of Pearson but do not receive any compensation for sales of the WAIS–IV in either country. Jianjun Zhu, Lawrence G. Weiss, and James A. Holdnack were employees of Pearson during the development of the WAIS–IV (Canadian and U.S. editions). James A. Holdnack is currently a consultant for Pearson. Donald H. Saklofske was an advisory panel member on the WAIS–IV standardization study in the United States and Canada.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

American Psychiatric Association (1968). *Diagnostic and Statistical Manual of Mental Disorders* (DSM-II), Second Edition, Washington, DC: Author.

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 Behavioural Science*, *20*, 154-166.

Bowden, S. C., Saklofske, D. H., & Weiss, L. G. (2010). 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.

Dash, U. N., Dennis, S. S., Mueller, H. H., Mancini, G. J., Snart, F. D., & Short, R. H. (1983). WISC-R subtest variability in a clinic-referred sample of Canadian children. *Canadian Journal of Behavioural Science*, *15*, 211-227.

Dumont, R., & Willis, J. O. (2002). *Severe discrepancy calculation by formula*. Retrieved from http://alpha.fdu.edu/psychology/severe_discrepancy_determination.htm (accessed January 15, 2015)

Ford, L., Swart, S., Negreiros, J., Lacroix, S., & McGrew, K. S. (2010). *Use of the Woodcock-Johnson III NU Tests of Cognitive Abilities and Tests of Achievement and Canadian Populations* (Woodcock-Johnson III Assessment Service Bulletin No. 12). Rolling Meadows, IL: Riverside.

Greenspan, S., & Woods, G. W. (2014). Intellectual disability as a disorder of reasoning and judgment: The gradual move away from intelligence quotient-ceilings. *Current Opinion in Psychiatry*, *27*, 110-116.

Hardman, S., & Oldridge, O. A. (1985). The predictive validity of the WISC-R on the Woodcock-Johnson Achievement Battery, the British Columbia Quick Individual Educational Test, and Teacher Ranking of Student Achievement. *Canadian Journal of School Psychology*, *1*, 31-38.

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., Silvestri, R., & Armstrong, I. T. (2015). Implications for educational classification and psychological diagnoses when using the Wechsler Adult Intelligence Scale-Fourth Edition with Canadian as opposed to American norms. *Journal of Psychoeducational Assessment*, *33*, 299-311.

Holmes, B. J. (1981). *Individually-administered intelligence tests: An application of anchor test norming and equating procedures in British Columbia* (Educational Research Institute of British Columbia Report No. 81, p. 11). Vancouver, Canada: Educational Research Institute of British Columbia.

Joint Advisory Committee of the Canadian Education Association. (1993). *Principles for fair student assessment practices for education in Canada*. Edmonton, Alberta: Author.

Kaufman, A. S., & Doppelt, J. E. (1976). Analysis of WISC-R standardization data in terms of the stratification variables. *Child Development*, *47*, 165-171.

Kunin, R., & Associates, Inc. (2012). *Economic impact of international education in Canada—An update*. Retrieved from http://www.international.gc.ca/education/report-rapport/economic-impact-economique/index.aspx?lang=eng

McDermott, P. A. (1995). Sex, race, class, and other demographics as explanations for children's ability and adjustment: A national appraisal. *Journal of School Psychology*, *33*, 75-91.

Pearson. (2010). *Wechsler Individual Achievement Test* (3rd ed.). Toronto, Ontario, Canada: Author.

Prairie Research Associates Inc. (2009). *Canada first: The 2009 survey of international students*. Retrieved from http://www.cbie-bcei.ca/media-centre/publications/research-reports/

Saklofske, D. H., Patterson, C. A., Gorsuch, R. L., & Tulsky, D. S. (2001). Discussing the guidelines for using the WAIS-III Canadian norms. In D. Wechsler (Ed.), *Wechsler Adult Intelligence Scale* (3rd Canadian ed., pp. 35-41). Toronto, Ontario, Canada: Harcourt Canada.

Saklofske, D. H., Reynolds, C. R., & Schwean, V. L. (Eds.). (2013). *Oxford handbook of child psychological assessment*. New York, NY: Oxford University Press.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications*. San Diego, CA: Author.

Schalock, R. L., Borthwick-Duffy, S. A., Bradley, V., Buntix, W. H. E., Coulter, M. D., Craig, E. M., . . . Yeager, M. H. (2010). *Intellectual disability: Definition, classification and systems of supports* (11th ed.). Washington, DC: American Association on Intellectual and Developmental Disabilities.

Stanovitch, K. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly 22*, 360-406.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1996). *Wechsler Intelligence Scale for Children* (3rd Canadian ed.). Toronto, Ontario, Canada: Harcourt Brace.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2001). *Wechsler Adult Intelligence Scale* (3rd Canadian ed.). Toronto, Ontario, Canada: Harcourt Canada.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). San Antonio, TX: Pearson.

Wechsler, D. (2004a). *Wechsler Intelligence Scale for Children* (4th Canadian ed.). Toronto, Ontario, Canada: Harcourt Assessment.

Wechsler, D. (2004b). *Wechsler Preschool and Primary Scale of Intelligence* (3rd Canadian ed.). Toronto, Ontario: Pearson Canada Assessment.

Wechsler, D. (2008a). *Wechsler Adult Intelligence Scale* (4th Canadian ed.). Toronto, Ontario, Canada: Pearson.

Wechsler, D. (2008b). *Wechsler Adult Intelligence Scale* (4th ed.). Bloomington, MN: Pearson.

Wechsler, D. (2012). *Wechsler Preschool and Primary Scale of Intelligence* (4th Canadian ed.). Toronto, Ontario, Canada: Pearson.

Wechsler, D. (2014a). *Wechsler Intelligence Scale for Children* (5th Canadian ed.). Toronto, Ontario, Canada: Pearson.

Wechsler, D. (2014b). *Wechsler Intelligence Scale for Children* (5th ed.). Bloomington, MN: Pearson.

Wilkins, C., Rolfhus, E., Weiss, L., & Zhu, J. J. (2005, April). *A new method for calibrating translated tests with small sample sizes*. Paper presented at the 2005 annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

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