

Gender Differential Item Functioning on the WISC-II:

Analysis of the Canadian Standardization Sample

Rebecca J. Gokiert and Kathryn L. Ricker

Centre for Research in Applied Measurement and Evaluation

University of Alberta

Introduction

The Wechsler Intelligence Scale for Children – Third Edition (WISC-III; Wechsler, 1991) is one of the most widely used psychological measures in the world. The WISC-III is an individually administered norm-referenced clinical instrument for assessing the intellectual ability of children aged six to sixteen. Ability on the WISC-III is measured as the Full-Scale (FSIQ) score: a compilation of 10 of the 13 subtests that combine to make up the test and is a measure of general intelligence, scholastic aptitude and readiness to master school curriculum (Sattler, 2001). Presently, within school systems in Canada, the WISC-III is among the most commonly used assessment tools for making high-stakes funding decisions for individuals with learning and behaviour exceptionalities. Given the extensive use of the WISC-III within the Canadian school system for selection and classification purposes, it is of the utmost importance to begin uncovering systematic performance differences among distinct groups of students and ask why these performance differences might occur.

While intelligence tests continue to be the most popular psychological measures given to children and adolescents, there has been a steady decline in bias investigations utilizing the most recent version of the WISC-III (Glutting, Oh, Ward & Ward, 2000). A recurring theme found in the few investigations conducted using the WISC-III suggests that males and females have different response strategies, which may affect overall IQ scores. Currently, males and females are not treated uniquely, and the same standard norms are applied to each group. Classical reviews of the literature on intelligence and achievement tests conclude that males outperform females on three types of abilities: general information, arithmetical reasoning and spatial ability. Conversely, females

outperform boys on measures of general verbal ability, spelling, grammar/language usage, rote memory and perceptual speed (Anastasi, 1958; Maccoby, 1966; Maccoby & Jacklin, 1974, cited in Feingold, 1993; Tyler, 1965).

Although there is some indication that males and females tend to differ in their academic achievement and subtests scores on measures of intelligence (Chen, 2000; Feingold, 1993; Maller, 2001), very few studies have investigated gender differences on the WISC-III. The majority of these inquiries have been conducted with summary statistical techniques focused at the subtest rather than item level (e.g., Chen, 2000; Slate, 1998, Willson, Nolan, Reynolds & Kamphaus, 1989). Moreover, the WISC-III manual and the Canadian supplement provide very little information regarding gender differences and the detection of DIF with Canadian students. To date, only one study has examined the WISC-III for gender differential item functioning (DIF), using the American standardization sample (Maller, 2001). DIF is a statistical method for analyzing differences in performance by assessing the probability that individuals of equal ability from each group would answer an item correctly (Hambleton, Swaminathan & Rogers, 1991). The advantage of using this type of technique is that it examines gender differences at the item versus test level. Once items have been identified as statistically significant for DIF, the next step is to determine whether this difference is due to bias or impact. An item is biased if the identified DIF is caused by a nuisance dimension irrelevant to what the test is designed to measure. This bias may result in the inconsistent selection and classification of students. Conversely, an item displays impact if a dimension that the test is intending to measure causes the detected DIF, because of knowledge and/or experience differences. This possibility begs the question of whether

unique gender norms are necessary on the WISC-III. The potential bias of specific test items towards certain groups has perhaps been the most highly charged issue surrounding testing today (Hambleton, et al., 1991). The Standards for Educational and Psychological Testing (1999) prescribe that tests must be free from bias to be fair.

Presently, there are no studies that have examined gender differences on the WISC-III using the Canadian national standardization sample. Generalizing findings from DIF analyses conducted with an American sample to a Canadian sample may not be meaningful given cultural differences that exist between these two countries.

Alternatively, the cultural differences between countries may reveal why gender differences occur on select items. Therefore, the purpose of the present paper is two-fold: 1) to explore gender DIF using the WISC-III in a Canadian context; and 2) to compare items that display DIF in the Canadian and American samples.

Method

The method used in the present paper was intended to replicate the procedures used by Maller (2001), because it is the only study thus far that has examined gender differences on the WISC-III using DIF detection procedures. The Canadian national standardization sample for the WISC III (N=1100) included children aged 6 through 16, split evenly by gender (550 females and 550 males) and is considered representative of the 1986 Canadian census with respect to age, gender, ethnic origin, parent education levels and geographic region. Detail regarding the sample is available in the Canadian supplement manual (Wechsler, 1996). Five of the 13 subtests were used in this study: Information (30 items), Vocabulary (30 items), Similarities (19 items), Comprehension (18 items), and Picture Completion (30 items). One additional subtest used by Maller,

Arithmetic, was excluded due to difficulties in parameter estimation because the items were speeded. Of the subtests included in this study, those items within each subtest share substantial similarities in those cognitive skills necessary to complete the items, and the format of the items. In essence, the items within a subtest are homogeneous.

All statistical analyses were conducted using the Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning (IRTLRDIF, Thissen, 2001). IRTLDRDIF is a software program for the computation of the statistics involved in IRT likelihood-ratio tests for DIF (Thissen, 2001). An exploratory approach to the statistical analysis was used. The suspect items were identified using an iterative purification approach. Iterative purification involves repeated analysis and removal of items that are flagged for gender DIF. Items are analysed in successive analyses until no further items are suspected to contain DIF. This subset of items, labeled the anchor test, are considered to be DIF-free and are used as a basis of comparison for the suspect items on the WISC-III. The IRTLDRDIF program fit the 2-parameter logistic model sequentially starting with the *a*-parameter (slope is set equal for the suspect item), then the *b*-parameter (both the *a*- and *b*- parameter are set equal for the suspect item). In the case of polytomously scored items, the model was fit for $m-1$ *b*-parameters (where m is the number of categories). After each iteration, items were flagged as suspect items and were removed from subsequent iterations. The criteria for flagging suspect items was an omnibus G^2 value greater than 3.84, which corresponds to the critical value when $\alpha = .05$ with one degree of freedom. The G^2 is a likelihood ratio statistic that evaluates the difference in the goodness of fit of the model between groups, and is distributed like a chi squared (X^2) statistic (Thissen, Steinberg & Gerrard, 1986). The iterative process continued until no

further items were flagged; these remaining items were labeled the purified anchor. Once the suspect subtest was identified, a final run was conducted where each individual suspect item was tested for DIF against the purified anchor. The G2 statistic for the a - and b - parameters were tested separately to determine if they were statistically significant at the $\alpha \leq .05$ level. Once the statistically significant parameters were found, the difference between the parameters for boys and girls was computed to determine the magnitude and direction of the differential item functioning. When an item displayed DIF for discrimination parameters (a - parameter), a positive difference indicated that the item was more discriminating for boys whereas a negative difference indicated that the item was more discriminating for girls. For difficulty (b - parameter), a positive difference indicated that the item was more difficult for boys (and thus favoured girls), and a negative difference indicate that the item was more difficult for girls (the items favour boys).

Results

Gender DIF in the Canadian Sample

Overall, 55 of 127, or 43.3% of items tested for gender DIF in the Canadian sample displayed significant DIF. Each subtest contained several items that showed DIF.

Information. Nine items displayed DIF out of 30 items. Items 4,7,9,10,15,17,22,24 and 25 displayed significant DIF for the b - parameter (Table 1). No items displayed significant DIF on the discrimination (a) parameter. Of the items displaying DIF, items 17 and 24 favoured boys, the remaining items favoured girls.

Vocabulary. Fourteen items displayed DIF out of 30 items. Items 2, 5, 6, 7, 10, 13, 15, 17, 18, 20, 23, 24, 25 and 28 were identified as displayed DIF using the omnibus

G^2 test (Table 2). Item 2 was not significant at the .05 level for either the *a*- or *b*-parameter. However, items 20 and 24 were significantly more discriminating for girls, while item 25 was more discriminating for boys. Items 5, 6, 7, 10, 13, 17, 18, 20, 23, and 24 favoured girls on the *b*- parameter, while items 15 and 28 favoured boys.

Similarities. Eleven items displayed DIF out of 19 items. Items 1, 2, 3, 5, 6, 7, 10, 15, 16, 17 and 18 displayed significant DIF (Table 3). Items 2 and 5 were more discriminating for boys, while items 10, 15 and 16 were more discriminating for girls. Items 5, 10, and 18 were easier for boys, while items 6, 7, 16 and 17 were easier for girls.

Comprehension. Ten items displayed DIF out of 18 items. Items 1, 2, 4, 5, 8, 9, 10, 13, 16, and 17 displayed DIF using the omnibus G^2 test statistic (Table 4). Of these items, item 4 was more discriminating for boys, while items 8 and 13 were more discriminating for girls. Items 2, 5, 8, 10, and 17 were more difficult for boys, while items 1, 4, 9, 13 and 16 were more difficult for girls.

Picture Completion. Twelve items displayed DIF out of 30 items. Items 8, 11, 12, 15, 16, 19, 20, 21, 24, 27, 28 and 30 displayed significant DIF (Table 5). Item 27 was significant on the omnibus G^2 test, but neither the *a*- or *b*- parameter displayed significant DIF. Items 8 and 12 were more discriminating for boys and girls, respectively. Additionally, item 12 favoured girls on difficulty, along with items 11, 15, 20 and 30. Items 16, 19, 21, 24, and 28 favoured boys.

Comparison with American sample results

Overall 20 of the 55 items (36.4%) with significant gender DIF in the Canadian sample also displayed significant DIF (favouring the same group) as in the American sample. The proportion of items that showed significant gender DIF in both samples

varied greatly from subtest to subtest. The Information and Picture Completion tests showed the most similarity between the Canadian and U.S. samples (7 of 9 and 8 of 12 Canadian DIF items, respectively). The Vocabulary, Comprehension and Similarities subtests had the fewest DIF items common to both samples (2 of 13, 2 of 10 and 1 of 11 Canadian DIF items, respectively).

In addition to the items that displayed significant DIF favouring the same group, two items from each of the Vocabulary and Similarities subtests were flagged for DIF in both samples, but for different reasons (meaning, the items favoured the opposite gender). For example, item 25 of the Vocabulary subtest was significantly more discriminating (*a*- parameter) for boys in the Canadian sample, while Maller found item 25 to be *easier* for boys in the American sample. Two items from the Comprehension subtest displayed DIF for the same parameter in both samples (*b*- parameter), favouring the opposite gender.

Discussion

A significant number of items in the Canadian sample showed gender differential item functioning. DIF items were present in all subtests, occurring across the level of difficulty (the easiest through hardest items) indicating the DIF was occurring at all levels of ability. Examining items that display DIF on the *b*-parameter, 29 items favoured girls, while 18 favour boys.

Two subtests (Information and Vocabulary) contained many more items favouring girls versus boys in terms of difficulty (7 of 9 and 10 of 12, respectively). Both of these subtests are verbal, and therefore girls would be expected to outperform boys on these tests based on previous research (Anastasi, 1958; Maccoby, 1966; Maccoby &

Jacklin, 1974, cited in Feingold, 1993; Tyler, 1965). However, these subtests require further substantive investigation to determine if performance differences that have been observed in the past are, in fact, due to impact, and not due to bias in the items that affects performance (i.e., better performance because of an irrelevant factor).

When comparing the Canadian and American samples for gender DIF approximately one third of items that display significant DIF display the same DIF across samples (20 of 55 DIF items). This result gives us greater confidence in asserting that these items contain some factor that contributes to differential item functioning between genders, and requires further investigation. The question of why the remaining items were not the same in both samples is another important question that must be answered. An even more complex question is *why* some items showed significant DIF in the opposite direction. These items might also be influenced by cultural differences between Canada and the United States.

The results of this study, alone and in combination with Maller's (2001) previous research, show that significant numbers of items on the WISC-III display gender DIF. Unlike Maller (2001), we argue that in the Canadian sample, these items are not distributed in such a way that the gender effects would balance out in the total score. All of these results, while exploratory in nature, suggest that systematic group differences occur on the WISC-III. Now, the next step is to understand from a substantive, cognitive perspective, why these gender differences are occurring.

References

- Anastasi, A. (1958). *Differential Psychology* (3rd ed.). New York: MacMillan.
- Chen, H. (2000). Gender differences in cognitive abilities: trends from age 6 to age 16 based on WISC-III standardization data for Taiwan. *Proc. Natl. Sci. Council. ROC(C)*, *10*, 201-216.
- Fiengold, A. (1993). Cognitive gender differences: a developmental perspective. *Sex Roles*, *29*, 91-112.
- Glutting, J.J., Oh, H., Ward, T., & Ward, S. (2000). Possible criterion-related bias of the WISC-III with a referral sample. *Journal of Psychoeducational Assessment*, *18*, 17-26.
- Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: SAGE Publications.
- Maccoby, E.E. (1966). Sex differences in intellectual functioning. In E.E. Maccoby (Ed.), *The Development of Sex Differences*. Stanford, CA: Stanford University Press.
- Maccoby, E.E., & Jacklin, C.N. (1974). *The Psychology of Sex Differences*. Stanford, CA: Stanford University Press.
- Maller, S.J. (2001). Differential item functioning in the WISC-III item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement*, *61*, 793-817.
- Sattler, J.M. (2001). *Assessment of Children: Cognitive Applications Fourth Edition*. La Mesa, California: Jerome M. Sattler, Publisher, Inc.

Slate, J.R., & Fawcett, J. (1996). Gender differences in Wechsler performance scores of school-aged children who are deaf or hard of hearing. *American Annals of the Deaf, 141*, 19-22.

Standards for Educational and Psychological Testing. (1999). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Thissen, D., Steinberg, L., & Gerrard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin, 99*, 118-128.

Thissen, D. (2001). IRTL RDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. University of North Carolina at Chapel Hill.

Tyler, L.E. (1965). *The Psychology of Human Differences* (3rd ed.). New York: Appleton.

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

Wechsler, D. (1996).

Willson, V.L., Nolan, R.F., Reynolds, C.R., & Kamphaus, R.W. (1989). Race and gender effects on item functioning on the Kaufman assessment battery for children.

Journal of School Psychology, 27, 289-296.

Table 1
Canadian vs American Differential Item Functioning (DIF) Results for the Information Subtest

Canadian					American				
Studied	G^2		G^2		Studied	G^2		G^2	
Item	$a = a$	$a_{boys} - a_{girls}$	$b = b$	$b_{boys} - b_{girls}$	Item	$a = a$	$a_{boys} - a_{girls}$	$b = b$	$b_{boys} - b_{girls}$
4	0.10	0.37	4.50	0.44	4	4.10	0.22	5.20	0.31
7	1.00	-0.56	4.10	0.11	7	0.30	-0.41	14.00	0.22
9	3.20	-1.23	10.90	0.16	9	0.00	0.02	14.80	0.25
10	0.10	-0.19	5.00	0.15	15	0.50	-0.22	31.00	0.32
15	2.30	-0.46	17.70	0.30	16	1.40	0.27	5.60	-0.18
17	0.30	-0.13	4.10	-0.18	17	0.80	0.31	12.70	-0.24
22	1.40	-0.61	36.30	0.53	19	0.00	0.04	4.60	0.09
24	0.40	0.39	12.80	-0.32	20	4.40	0.78	9.30	-0.25
25	0.60	0.55	10.30	0.19	21	3.10	1.29	6.50	-0.25
					22	2.40	-1.30	33.90	0.42
					24	0.50	0.36	21.00	-0.37
					27	0.50	0.37	6.70	-0.26
					29	0.30	-0.40	4.60	-0.11
					30	0.10	-0.25	15.00	-0.55

Note. When an item displayed DIF for discrimination parameters, a positive difference indicated that the item was more discriminating for boys, whereas a negative difference indicated that the item was more discriminating for girls. In terms of item difficulty, a positive difference indicated that the item was more difficult for boys, and a negative difference indicated that the item was more difficult for girls.

Table 2
Canadian vs American Differential Item Functioning (DIF) Results for the Vocabulary Subtest

Canadian						American						
Studied	G^2		G^2			Studied	G^2		G^2		G^2	
Item	$a = a$	$a_{boys} - a_{girls}$	$b = b$	$b_{boys} - b_{girls}$	$b_{boys} - b_{girls}$	Item	$a = a$	$a_{boys} - a_{girls}$	$b_1 = b_1$	$b_{boys} - b_{girls}$	$b_2 = b_2$	$b_{boys} - b_{girls}$
5	0.40	0.26	7.90	0.76	0.30	12	0.00	0.08	10.00	-0.15	15.90	-0.18
6	1.10	-0.26	16.10	0.55	0.19	13	0.30	-0.11	2.70	0.08	5.20	0.18
7	0.50	0.17	6.40	0.22	0.37	15	6.30	0.74	3.20	0.09	3.20	0.09
10	1.90	0.31	4.60	0.48	0.42	24	12.10	-1.59	4.60	0.18	6.90	0.38
13	0.60	0.18	4.60	0.14	0.16	25	0.50	0.30	1.70	-0.12	9.90	-0.33
15	0.30	0.23	5.40	0.12	-0.15	26	0.00	0.06	8.30	-0.23	6.90	-0.27
17	0.20	-0.17	6.60	0.00	0.11	27	3.70	-1.32	7.50	0.05	10.50	0.02
18	0.10	0.2	6.30	0.14	0.14							
20	5.50	-0.62	5.00	0.10	0.15							
23	1.80	-1.38	16.10	0.28	0.22							
24	6.80	-2.26	10.80	0.09	-0.07							
25	4.80	1.29	3.70	-0.18	0.00							
28	1.00	1.22	5.40	-0.42	-0.26							

Note. When an item displayed DIF for discrimination parameters, a positive difference indicated that the item was more discriminating for boys, whereas a negative difference indicated that the item was more discriminating for girls. In terms of item difficulty, a positive difference indicated that the item was more difficult for boys, and a negative difference indicate that the item was more difficult for girls.

Table 3
Canadian vs American Differential Item Functioning (DIF) Results for the Comprehension Subtest

Canadian						American							
Studied	G^2		G^2			Studied	G^2		G^2		G^2		
Item	$a = a$	$a_{boys} - a_{girls}$	$b = b$	$b_{boys} - b_{girls}$	$b_{boys} - b_{girls}$	Item	$a = a$	$a_{boys} - a_{girls}$	$b = b$	$b_{boys} - b_{girls}$	$b = b$	$b_{boys} - b_{girls}$	
1	1.90	0.44	3.10	2.38	0.95	1	1.00	0.38	0.00	0.39	5.70	0.06	
2	1.40	-0.19	5.30	-2.53	0.05	3	8.40	-0.93	1.50	-1.43	0.00	-0.77	
4	3.00	0.51	6.30	0.65	0.75	4	6.50	-1.79	3.20	-0.41	1.10	-0.18	
5	0.10	-0.05	5.30	0.50	-0.32	6	0.10	0.04	0.00	0.06	4.80	0.69	
8	8.50	-0.46	2.70	-0.50	-0.31	7	2.00	-0.19	5.20	0.02	0.40	0.15	
9	0.00	0.02	6.00	0.21	0.10	10	1.00	-0.23	2.70	-0.18	4.90	0.15	
10	1.30	0.25	4.30	-0.14	-0.10	13	1.50	-0.47	5.10	0.11	3.10	0.14	
13	6.30	-1.10	2.90	0.07	0.09								
16	2.20	0.64	7.20	0.14	0.05								
17	0.10	-0.17	10.30	-0.18	-0.11								

Note. When an item displayed DIF for discrimination parameters, a positive difference indicated that the item was more discriminating for boys, whereas a negative difference indicated that the item was more discriminating for girls. In terms of item difficulty, a positive difference indicated that the item was more difficult for boys, and a negative difference indicate that the item was more difficult for girls.

Table 4
Canadian vs American Differential Item Functioning (DIF) Results for the Picture Completion Subtest

Canadian					American				
Studied	G^2		G^2		Studied	G^2		G^2	
Item	$a = a$	$a_{boys} - a_{girls}$	$b = b$	$b_{boys} - b_{girls}$	Item	$a = a$	$a_{boys} - a_{girls}$	$b = b$	$b_{boys} - b_{girls}$
8	6.20	1.09	3.60	0.72	5	2.10	-0.94	6.20	-0.05
11	0.70	-0.24	6.00	0.18	11	1.60	-0.34	9.10	0.10
12	5.70	-1.09	19.50	0.21	12	0.40	-0.27	41.80	0.43
15	0.20	-0.15	20.50	0.44	14	5.20	-0.57	0.10	-0.05
16	1.10	0.28	4.40	-0.13	15	0.00	-0.07	8.10	0.18
19	0.70	0.21	9.10	-0.27	16	0.20	0.18	22.30	-0.28
20	0.20	-0.15	4.70	0.16	18	8.00	1.06	4.50	-0.15
21	0.60	-0.17	5.70	-0.29	19	2.60	0.43	10.70	-0.18
24	0.00	-0.07	25.20	-0.48	21	0.10	0.06	16.40	-0.36
28	1.60	-0.61	4.80	-0.09	22	1.30	0.26	6.50	-0.25
30	3.40	-0.86	7.40	1.25	23	0.10	0.10	19.50	-0.26
					24	1.40	0.32	37.90	-0.52
					25	0.10	-0.10	8.80	-0.21
					28	0.10	0.07	17.40	-0.43

Note. When an item displayed DIF for discrimination parameters, a positive difference indicated that the item was more discriminating for boys, whereas a negative difference indicated that the item was more discriminating for girls. In terms of item difficulty, a positive difference indicated that the item was more difficult for boys, and a negative difference indicate that the item was more difficult for girls.

Table 5
Canadian vs American Differential Item Functioning (DIF) Results for the Similarities Subtest

Canadian						American						
Studied	G^2		G^2			Studied	G^2		G^2		G^2	
Item	$a = a$	$a_{boys} - a_{girls}$	$b = b$	$b_{boys} - b_{girls}$	$b_{boys} - b_{girls}$	Item	$a = a$	$a_{boys} - a_{girls}$	$b_1 = b_1$	$b_{boys} - b_{girls}$	$b_2 = b_2$	$b_{boys} - b_{girls}$
1	11.00	-2.78	0.70	-0.50	--	7	8.60	-0.79	1.00	-0.25	0.20	0.06
2	4.20	3.41	0.20	0.11	--	8	4.10	0.76	11.20	0.25	11.50	-0.20
3	0.40	4.55	4.00	-0.45	--	10	3.70	-0.90	6.40	0.08	5.00	0.08
5	10.10	1.30	5.20	1.14	--	13	3.90	-0.56	0.20	0.03	1.50	0.12
6	1.80	4.22	2.10	-0.34	0.03	15	1.10	0.36	4.00	-0.13	8.30	-0.23
7	1.60	-0.41	8.90	-0.22	0.16	16	0.00	0.08	6.60	-0.15	0.90	-0.12
10	4.20	-1.26	7.50	0.11	0.10							
15	5.20	-0.99	1.50	0.01	0.12							
16	5.90	-0.89	6.70	-0.14	0.28							
17	0.20	0.20	4.50	-0.17	-2.30							
18	1.60	-0.50	8.70	0.20	0.65							

Note. When an item displayed DIF for discrimination parameters, a positive difference indicated that the item was more discriminating for boys, whereas a negative difference indicated that the item was more discriminating for girls. In terms of item difficulty, a positive difference indicated that the item was more difficult for boys, and a negative difference indicate that the item was more difficult for girls.