Running head: COMPARISON OF WISC-III REGRESSION LINES ACROSS GENDER AND REGION

Comparison of WISC-III Regression Lines across Gender and Region for School Achievement Prediction in Taiwan

Hsin-Yi Chen

Department of Special Education

National Hualien Teachers College

Hualien, Taiwan, R.O.C.

Comparison of WISC-III Regression Lines

Abstract

This study compared the regression lines for the prediction of school achievement by the Wechsler Intelligence Scale for Children-Third Edition (WISC-III) Full Scale IQ across gender and region through the Potthoff analysis, which allows a simultaneous test of intercepts and slopes across groups. Non-standardized teacher assigned classroom grade was used as the index for school achievement. Data based on a total of 1,100 children age 6 to 16 in the WISC-III Taiwan standardization sample was analyzed. The important findings were: (1) Regression lines across regions (West vs. East., and North, Central, South, vs. East) did not differ significantly. It generally supported the fairness of using a common regression line in the prediction of school achievement scores across different regions in Taiwan. (2) Comparison of regression lines across gender showed a same slope but different intercepts, though the effect of the intercept difference was considered as “small”. Both the non-standardized achievement index and the relationship between gender role and teacher’s grading were discussed as possible explanations. Suggestions for future studies were also discussed.

Key Words: WISC-III, IQ, Achievement, Test Bias, Regression, Gender, Region

Comparison of WISC-III Regression Lines across Gender and Region for School Achievement Prediction in Taiwan

It is well known that Intelligence and achievement correlate positively with each other (Fisher, 1995; Lassiter, 1995; Lavin, 1996; Slate, 1994; Smith et al, 1995). Intellectual abilities are usually considered as potential capacities. Achievement, on the contrary, stands for the current performance. In the real setting, intellectual abilities have long been used as the predictors for achievement (Glutting et al., 1997; Kaplan, 1996).

Among all the intellectual indexes, the full scale IQ (FSIQ) in the Wechsler Intelligence Scale for Children-Third Edition (WISC-III) has been widely used for achievement prediction (Reynolds & Hartlage, 1979; Weiss, Prifitera, & Roid, 1993) for many reasons. First of all, Wechsler's intelligence test is one of the most popular IQ tests universally. Secondly, the FSIQ was originally designed to measure the global ability,

which many researchers in this field accept as a universal description of intelligence, that is, the g ability. The FSIQ thus is usually being considered as a first step information in beginning to understand a child's overall functioning. Finally, the FSIQ is the index with the highest reliability among all WISC-III measures.

In 1968, the American Psychological Association (APA) Board of Scientific Affairs appointed a committee to investigate the issue of fairness in using psychological and educational tests. The committee considered several previous viewpoints (Cleary, 1968; Potthoff, 1966) and recommended a definition of test bias in 1975, which not only considered the usually cited content validity and construct validity, but also focused exclusively on the issue of predictive validity. According to them, "A test is considered fair for a particular use if the inference drawn from test score is made with the smallest feasible random error and if there is no constant error in the inference as a function of membership in a particular group." (Cleary, Humphreys, Kendrick, & Wesman, 1975, p.25). This definition is later known as the regression definition of test bias (Reynolds, 1982).

In 1975, the U.S. public law 94-142: The education for all handicapped

children act of 1975(Federal Register, 1977) mandated that tests can not be developed and normed based on majority group and consequently biased against minorities. In 1985, APA, the American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME) placed joint guidelines on this fairness issue (American Psychological Association, 1985). In 1991, "The individuals with disabilities education act (IDEA)" was regulated (Federal Register, 1991). All these above actions showed that the test fairness across groups has become a central issue with increasing concerns. Various studies and methodological reviews have also emerged (e.g. Berk, 1982; Reynolds & Kaiser, 1990).

"Test bias" is actually a term with multifaceted sources. Besides possible bias due to content or construct inappropriate, the other source could be from the prediction process, which strongly related to the predictive validity of the test. As Reynolds (1982) stated, " Since our definition of bias in predictive validity requires errors in prediction to be independent of group membership, the regression line formed for any pair of variables must be the same for each group for whom prediction is to be made. Whenever the slope or the intercept differs significantly

across groups, there is bias in prediction if one attempts to use a regression equation based on the combined groups." To restate this concern, test bias is said to occur if one only regression line is used for practical prediction, while regression lines for subgroups differ significantly.

In Taiwan, the WISC-III Taiwan version (Wechsler, 1997) is a newly adapted instrument, the efforts made in the developing process, such as the representative of the standardization sample, the carefulness in item selections, ...etc, all provided the evidence for equivalency of internal psychometric properties across subgroups. (Wechsler, 1991, 1997). However, the predictive validity under the regression definition has not been examined for this newly developed version.

The purpose of this present study was to test the differential prediction of school achievement by the WISC-III FSIQ across various subgroups. Traditionally, gender difference is one important issue in studies of group difference (Willerman, 1979). In Taiwan, the degree of economic and educational development, also the combinations of racial populations, both are known to differ between geographic regions (Fu, 1994). For these reasons, three cross-group comparisons were tested in this current study: (1) Gender difference: boys vs. girls. (2) Geographic

region difference: North, Central, South, and East sides of Taiwan. Each region was compared to the other three. (3) Geographic region difference: West (combined all North, Central and South parts as one group) vs. East sides of Taiwan.

Method

Subjects

The data analyzed was based on the 1,100 standardization cases reported in the WISC-III Taiwan edition manual (Wechsler, 1997). This sample was selected based on a stratified random sampling plan for ensuring a close match between this sample and the Taiwan children population reported by the Taiwan Bureau of the Census in 1992. The following sections present the characteristics of this nationally representative sample:

(1)Geographic Region:

There were four major geographic regions specified in the Census reports: North, Central, South, and East. Based on the population stratification, this sample selected children in accordance with the proportions of children living in each region: North (42%), Central

(26%), South (28%) and East (4%).

(2)Age:

This sample included 100 children in each of 11 age groups ranging from 6 through 16 years old. Within each age group, subjects were carefully selected to cover the whole year ranging from 0 month 0 days to 11 month 30 days.

(3)Gender:

The sample included roughly 50 boys and 50 girls in each age group, with a little fluctuation for each age group.

(4)Parents Education Level:

There were five education categories specified in the Census reports. This sample selected proportions of subjects based on parent education level: College or above (10.5%), Technical School (12.5%), Senior High-School (35.3%), Junior High-School (22.9), and Elementary School or below (18.7%).

(5)Urban vs. Rural District:

Subjects were selected from schools in both urban and rural district based on population proportions.

(6)Ethnicity:

Considering the representative of aborigines in Taiwan, at least 2 children in each of the 11 age groups were aborigines from East Taiwan.

For each subject, the FSIQ was calculated and school achievement based on teacher-assigned classroom grade (range from one to three) was gathered. This achievement score is non-standardized since the differences in grading among teachers were not controlled.

Technique

To compare the regression lines across groups, some researchers suggested methods for testing intercepts and slopes (regression coefficients) across groups separately (Gulliksen & Wilks, 1950). This method is considered to have the drawback of inflating the type one error by testing two values separately (Reynolds, 1982). Potthoff (1966, PP 18-21) provided the other technique, which allows one simultaneous test of the intercepts and slopes across groups with a single F ratio (Potthoff test value F1). If a significant F results, it means that the homogeneity of regression across groups does not occur, slopes and intercepts may then be assessed separately (Potthoff test value F2 and F3) to see which value differs, or whether both are different. On the other hand, if a non

significant F results, it means the regression lines across groups are the same, thus no more tests on slopes and intercepts will be necessary.

Procedure

Before tests being conducted, the type one error rate was decided to be 0.01 for reasons of multiple comparisons. For each one of the three comparisons, the Pothoff's test value F1 was first evaluated to check the homogeneity between subgroup regression lines, where the dependent variable is the school achievement and the independent variable is the FSIQ. If a significant F1 value results, the slopes and intercepts are tested further by Pothoff's test value F2 and F3. If any of these two values was significant, the degree of difference was evaluated by Cohen's effect size approach (1988).

Results

Descriptive statistics for each subgroup on both FSIQ and school achievement measures are reported in Table 1. Also presented in this table is the correlation between FSIQ and school achievement. All correlations are significant at the α=0.01 level, though these correlation values are somewhat smaller than what is usually obtained between intelligence and academic performance measures, 0.5 (Brody, 1992). The restriction of

range, sample size effect, and use of non-standardized teacher grading could be possible reasons.

Table 1

Descriptive Statistics and Correlations of WISC-III FSIQ and School Achievement Across Subgroups

			WISC-III Full scale IQ		School Achievement		
		n	*M*	*SD*	*M*	*SD*	*r*
Gender							0.44*
	Boys	548	101.99	16.18	2.02	0.79	0.51*
	Girls	550	98.94	14.11	2.07	0.77	0.37*
Region							0.43*
	West	1003	101.03	14.51	2.06	0.78	0.43*
	North	440	103.68	15.08	2.10	0.77	0.37*
	Central	257	97.02	13.10	2.02	0.79	0.43*
	South	306	100.58	14.04	2.03	0.80	0.51*
	East	54	92.02	14.86	2.00	0.80	0.49*

Note. * $p<0.01$

Table 2 presents the result of Pothoff tests for all three comparisons. For comparisons across regions, F1 values are nonsignificant ($p>0.01$). However, a significant F1 effect was observed for the prediction of

teacher assigned grade in the boys and girls comparison. Consequent Pothoff value F2 was not significant, and Pothoff value F3 was significant. This result suggests a same slope but different intercepts between gender group regression lines.

Table 2

Pothoff's F-tests for Regression Lines Across Various Group Comparisons

Comparisons	F1(df)	F2(df)	F3(df)
Gender			
Boys vs. Girls	5.25* (2, 1094)	2.65 (1, 1094)	7.86* (1, 1095)
Region			
North, Central, South, vs. East	2.50 (6, 1049)	2.6 (3, 1049)	2.4 (3, 1052)
West vs. East	1.62 (2, 1053)	0.35 (1, 1053)	2.89 (1, 1054)

Note. * $p<0.01$

Regression lines based on the boys, girls, and the combined group were shown in Figure 1. The same slope was estimated by the combined group slope (0.023), and the intercept for boys is –0.515, for girls is 0.059, and for combined group is –0.217. Based on this figure, the difference

between genders on predicted classroom grade is the same across various FSIQ. This is due to the same slope result. Thus, there is no differential predictive difference between different FSIQ points to be worried.

It is clearly detected from the figure also that when the combined group regression line was used as the one for both gender, the predicted classroom grade for girls tend to be underestimated and the predicted score for boys tend to be overestimated. Girls tend to be underestimated by 0.276 achievement point when the combined group regression line was applied. The standardized difference between predicted grade based on the girls' group and combined group is roughly 0.39 standard error of measurement units (using SEM=0.70 for predicting teacher assigned classroom grade by the FSIQ in the combined sample with 1098 cases). Based on the subjective convention offered by Cohen (1988, p.24-27), the effect size of this difference is considered as "small".

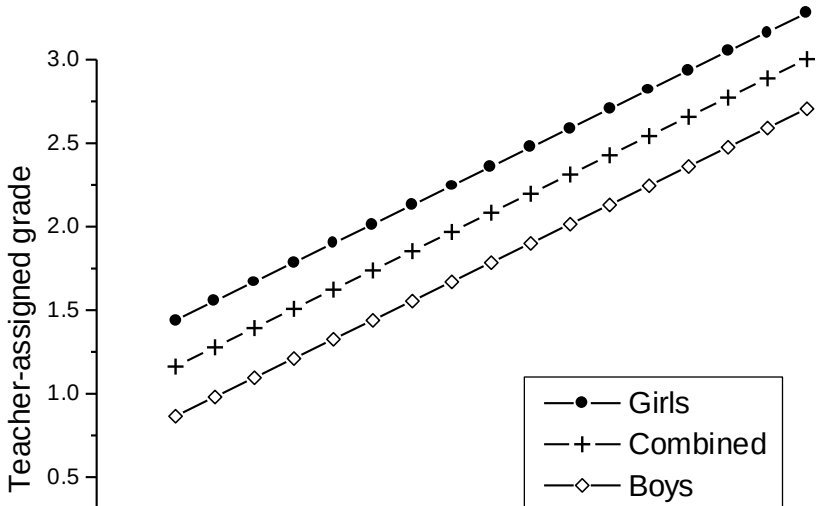

Figure 1. Regression lines (same slope, different intercepts) for boys (*n*=548), girls (*n*=550), and combined group (*n*=1098). The same slope was estimated by using the combined group slope.

Discussion and Suggestion

This study is important for two reasons. First, it provided the information on how to maintain predictive fairness for the newly released WISC-III across subgroups in Taiwan. Second, the data analyzed is the standardization sample, which well represents the Taiwan children population from age 6 to age 16.

Results showed that when the non-standardized teacher-assigned classroom grade was used as the index of school achievement, using one regression line to predict achievement is fair across regions. That is, using one only prediction line is fair for children from either West (North, Central or South), or East side of Taiwan. Given the general concerns of

the unbalanced development between Taiwan regions, this result deserves attention.

While testing the comparability of the regression lines across gender, results showed a same slope but different intercepts. It suggested that when using only one prediction line to predict the school grade for both sexes, there might be a bias against toward girls, for their predicted achievement score tend to be systematically underestimated by predicted score from that combined group regression line. One extremely important explanation needs to be clarified again, this result of differential regression lines across gender should not lead readers jump to the immature conclusion that“ the WISC-III is biased”. Instead, this result suggests that, when predicting the non-standardized teacher-assigned classroom grade, using the one combined group regression line for both boys and girls may not be appropriate. Separate lines are needed for predictions.

There are several limitations of this study. First, measures of achievement varies, and the one used in this current study was the non-standardized teacher assigned classroom grade (ranges from one to three). Weiss, Prifitera, and Roid (1993) reported similar finding that

when predicting non standardized teacher-assigned English grade by WISC-III FSIQ across gender, different intercept was detected with a same slope. However, when using score from a nationally standardized English test as the predicted score, no predictive difference was found. Hsieh (1992) investigated the academic achievement scores of 795 junior high school students in Taipei across gender, what she found was that girls outperform boys for the teacher-assigned school grade as usually noticed, however, boys outperform girls on the senior high school entrance examine. These findings suggest the possibility that gender role (boys vs. girls) given by social cultural influence may play a bigger role on teacher's judgement, rather than on the standardized achievement score. These two criterion (teachers' judgement vs. standardized achievement score) are without doubt both measuring " school achievement", but may touch somewhat different aspects of it.

The other limitation of this study is the range of teacher assigned school achievement, from one to three. This small range limits potential achievement variability, and thus could lead to a smaller correlation and somewhat different regression lines.

Since the "predictive method" has been widely accepted as one of the

ways of diagnosing children with learning disability. Ensure the fairness of regression line using is extremely important. In this case, for example, the predicted achievement score based on the common regression line tend to overestimate boys' achievement. It then will lead to a bigger difference between the predicted and actual achievement, finally, it may results in having more boys being diagnosed as the "learning disabilities".

According to author, this study plays the role of "a starting alert" for people in this field. More and more validation investigations on the issue of prediction fairness are eagerly needed in order to provide a much clear picture. Future studies are suggested to work on the followings. First, using achievement score based on the standardized achievement test may provide a purer estimate of achievement, which is less influenced by the social-cultural affected gender roles. Second, selecting variables with larger variance. Third, Detecting predictive fairness across racial groups, like comparing the predictive fairness between aborigines and non-aborigines; or checking other higher order interaction effects.

References

American Psychological Association (1985). Standards for educational and psychological testing. Washington, DC: Author.

Berk, R. A. (Eds.). (1982). Handbook of methods for detecting test bias. Baltimore: Johns Hopkins University Press.

Brody, N. (1992). Intelligence (2nd ed.). San Diego, CA: Academic Press.

Cleary,T.A., Humphreys, L.G., Kendrick, S.A., & Wesman, A. (1975). Educational uses of tests with disadvantaged students. American Psychologist, 30, 15-41.

Cohen, J. (1988). Statistical power analyses for the behavior sciences. Hillsdale, NJ: Erlbaum.

Federal Register (August 23, 1977). Regulations implementing the Education for all Handicapped Children Act of 1975, pp. 42474-42518.

Federal Register (August 19, 1991). Regulations for implementing the Individuals with Disabilities Education Act, Vol. 56, pp. 41266-41275.

Fisher, J. L. (1995). Relationship of intelligence quotients to academic achievement in the elementary grades. ERIC Document Reproduction Service No. ED 388 428.

Fu, Y, C. (1994). Distinctive ethnic group position in eastern Taiwan. National Taiwan University Journal of Sociology, 23, 191-216.

Glutting, J. R., Youngstrom, E. A., Ward, T., Ward, S., & Hale, R. L.(1997). Incremental efficiency of WISC-III factor scores in predicting achievement: what do they tell us? Psychological Assessment, 9(3), 295-301.

Gridley, B. E. & Roid, G. H. (1998). The use if the WISC-III with Achievement tests. In A. Prifitera & D. Saklofske (Eds.), WISC-III Clinical Use and Interpretation Scientist-Practitioner Perspectives (pp. 249-288). San Diego, CA: Academic Press.

Gulliksen, H., & Wilks, S. S. (1950). Regression tests for several samples. Psychometrika, 15, 91-114.

Hsieh, H, C. (1992). Gender difference in educational opportunity in Taiwan: Two Taipei junior high schools. Proceedings of the National Science Council. Republic of China. Part C: Humanities and Social Sciences, 2(2), 179-201.

Kaplan, C. (1996). Predictive validity of the WPPSI-R: a four year follow-up study. Psychology in the Schools,33(3), 211-220.

Lassiter, K. S. (1995). The relationship between young children's academic

achievement and measure of intelligence. Psychology in the Schools,32(3), 170-177.

Lavin, C. (1996). The relationship between the Wechsler Intelligence Scale for children-third edition and the Kaufman test of education achievement. Psychology in the Schools, 33(2), 119-123.

Potthoff, R. F. (1966). Statistical aspects of the problem of biases in psychological tests (Institute of Statistics Mimeo Series No. 479). Chapel Hill: Department of Statistics, University of North Carolina.

Reynolds, C. R. (1982). Methods for detecting construct and prediction bias. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp.199-259). Baltimore: Johns Hopkins University Press.

Reynolds, C. R., & Hartlage, L. (1979). Comparison of WISC and WISC-R regression lines for academic prediction with black and with white referred children. Journal of Consulting and Clinical Psychology, 47(3), 589-591.

Reynolds,C. R., & Kaiser, S. M. (1990). Test bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), The handbook of school psychology (pp. 487-525). New York: John Wiley.

Slate, J. R. (1994). WISC-III correlations with the WIAT. Psychology in the Schools, 31(4), 278-285.

Smith, T.D., and others (1995). The relationship between the WISC-III and the WRAT3 in a sample of rural referred children. Psychology in the Schools, 32(4), 291-295.

Wechsler, D. (1944). The measurement of adult intelligence (3rd ed.). Baltimore: Williams & Wilkins.

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

Wechsler, D. (1997). Manual for the Wechsler Intelligence Scale for Children - Third Edition. Taiwan Edition. Taipei, Taiwan: The Chinese Behavior Science Corporation.

Weiss, L.G., Prifitera, A., & Roid, G. (1993). The WISC-III and the fairness of predicting achievement across ethnic and gender groups. Journal of Psychoeducational Assessment Monograph Series. Advances in Psychological Assessment: Wechsler Intelligence Scale for Children-Third Edition, 35-42.

Willermane, L. (1979). The Psychology of Individual and Group Differences. San Francisco, CA: W.H. Freeman and Company.

Comparison of WISC-III Regression Lines

以智力預測學業成就：不同性別與地理區域內迴歸線之比較研究

陳心怡

國立花蓮師範學院 特殊教育學系

摘要

本研究之目的在探討魏氏兒童智力量表(WISC-III)之全量表智商分數(FSIQ)在預測學業成就上之測驗偏頗性。所採用之樣本是臺灣地區 WISC-III 標準化常模中 1,100 位六至十六歲兒童，每位兒童之學業成就是以其班級老師給與之學業成績爲指標。研究者以 Potthoff 方法同時考驗依不同性別與不同地理區域建立之各組迴歸線在斜率與截距上之一致性。迴歸線之一致性與測驗之預測效度有高度關聯:因爲在不同組別迴歸線有差異之情況下，如果仍然以一條單一迴歸線來進行成就預測，測驗偏頗性便會產生並對某些組別之兒童造成實際上不公平之影響。研究結果發現: (1) 根據兩組臺灣地理區域之比較 (一爲北部、中部、南部、與東部之比較，另一爲西部與東部之比較)，各組迴歸線沒有顯著差異。此結果顯示用合併各組所建立之單一迴歸線來以智力預測學業成就並不會造成測驗偏頗性。亦即不會對來自任一地理區域內之學童造成不公平之影響。(2)男女學童組內之迴歸線有一樣的斜率，但女生組之截距卻顯著偏高。初步建議以不同迴歸線對男女學童進行預測較爲適當。而在此同時，需要更多以標準化成就測驗分數爲效標之效度研究，以期對測驗偏頗性研究有更正確之瞭解。

關鍵詞: 魏氏兒童智力量表、智力、成就、測驗偏頗性、迴歸、性別、地域。