

Paper code “AGHO5116”

Estimating the Hausman test for Rasch with poorly fitting items

Kingsley E. Agho¹ and James A. Athanasou²

¹ Centre for Clinical Epidemiology and Biostatistics, University of Newcastle,

² Faculty of Education, University of Technology, Sydney

Abstract

In this study, an assessment that was difficult for a sample was used as a demonstration of the bootstrap and simulation method for estimating the Hausman test for Rasch analysis when the items are also poorly fitted. A 10-item dichotomously scored test of numerical reasoning was administered to 200 (120 male, 80 female) high school pupils in Nigeria. An initial analysis using RUMM showed that the fit of the items to the Rasch model was poor and 1000 bootstrap replicates of the sample were generated. This paper reports results from the parametric, simulation and bootstrap method for estimating the Hausman test for the Rasch model. The main findings were that the simulation and bootstrap method for estimating the Hausman test for Rasch were statistically better than the parametric method and there was no need to eliminate poorly fitted items as suggested previously in the literature.

Key words: Rasch model, Hausman test

Corresponding author, Tel.: 61249236901

E-mail address: kingsley.agho@newcastle.edu.au (K. Agho)

Since, the introduction of the Rasch (1960) measurement model to educational and psychological testing, several test for determining model fit in Rasch has been proposed. The two most commonly used for evaluating goodness of fit test for Rasch are chi-square fit statistics (Wright & Panchapakassen, 1969; Linacre & Wright, 1994) and the conditional maximum likelihood method (Andersen, 1973a,b). Other model selections for Rasch are the Wald test, the Lagrange Multiplier test (Fischer & Molenaar, 1995) and the Hausman test (Hausman, 1978; Weesie, 1999).

There are substantive and technical overlaps between the purpose of judging fit over all the available data and the purpose of isolating misfitting (Traub & Wolfe, 1981). According to Traub & Wolfe (1981), a researcher who wishes to assess model fit would certainly want to do so using the responses of only those people who took the task of answering the items seriously. What is obvious is that the assessments of model fit cannot be taken seriously if persons and items are culled on the basis of preliminary applications of the latent trait analysis before the fit of the model to the remaining data is tested (Traub & Wolfe, 1981). Such a prior analysis misguided the logic of statistical tests on which we base claims of model fit (Traub & Wolfe, 1981). On the other hand, the conditional maximum likelihood approach employed by Andersen (1973a,b) to determine fit statistics are flawed because the Maximum likelihood theory is not applicable to joint estimation and Andersen's likelihood ratio is distributed as a chi-square only if the number of persons in each sub-sample of the total sample is reasonable large (Traub & Wolfe, 1981).

In recent studies, nonparametric methods (including the bootstrap) have been used to deal with items that did not fit the Rasch model well. For example, Douglas and Cohen (2001) used the semiparametric approach to be applied simultaneously to an entire set of items. The semiparametric procedure puts items into two classes. They estimated those that fit well in the parametric model by using a method of unidimensional parametric item response model item calibration and those items that did not fit well were estimated using the nonparametric techniques. This method is semiparametric because some of the items were fitted parametrically while others were fitted nonparametrically. This method has

two advantages, one statistical and the other substantive. The approach is beneficial statistically to item response models (including Rasch) because the data would fit better than with a totally parametric approach (Stout, 2001). The substantive advantage is that one is able to test that the items are poorly fitted.

The purpose of this paper is to demonstrate an application of the bootstrap and simulation method to estimate the Hausman test for Rasch as a possible application to the problem of model selections when the data did not fit the Rasch model. Throughout the paper, we will discuss two nonparametric methods for estimating the Hausman test and we stress that the proposed methods are easy to implement in any educational testing context. All that is needed in this procedure is the program for simulating or bootstrapping data, a program for estimating the conditional Rasch model and a program for estimating the Hausman test. In the next two sections, the parametric Hausman test and its bootstrap procedure in Rasch will be discussed.

This paper builds upon the work of Andersen, (1973a,b), Weesie, (1999), Fisher and Molenaar, (1995), Douglas and Cohen, (2001) to develop the bootstrap method for estimating the Hausman test for Rasch. In this study, we selected an assessment that was difficult for a sample (120 male, 80 female high school pupils in Nigeria) because the test was initially designed for higher ability students sitting for the West African Examination Council Senior Secondary School leaving certificate syllabus and we used it as a practical demonstration for using the bootstrap method to estimate the Hausman test for Rasch when the items are poorly fitted. In this paper we first describe the Hausman test for Rasch and then we describe the procedures for using the bootstrap method to estimate the Hausman test for Rasch. An application of the parametric and bootstrap methods to estimate the Hausman test for Rasch is presented using the mathematics test as an example.

The Hausman test for Rasch

The Hausman test is closely related to the Likelihood Ratio test, Wald test and Lagrange Multiplier test. The Hausman test is based on the difference between two difficulty parameter estimates. For example, the Hausman test will enable researchers to compare the item difficulty parameters obtained from the full sample with the item difficulty parameters obtained say a female sub-sample or male sub-sample. The null hypothesis of the Hausman test is the same as the Likelihood Ratio test, Wald test and Lagrange Multiplier test and they have the same asymptotic power for local alternatives (Fisher & Molenaar, 1995; Weesie, 1999).

In this study, the generic term ‘item parameter’ will be used to refer to component δ to explore the Rasch model. Consider the following response data situation, where i represents dichotomous responses given to j items. Let X_{ij} be the binary or dichotomous (1,0) response for person i ($i = 1, \dots, N$) and item j ($j = 1, \dots, n$), where 1 denotes a correct response and 0 denotes an incorrect response. Let $P_{ij} = P(X_{ij} = 1)$ and $Q_{ij} = 1 - P_{ij} = P(X_{ij} = 0)$. Furthermore, let X_{ij} denote the full data which has been subdivided into X_{ijF} and X_{ijM} which denote the female (F) and male (M) samples. The simplest and the most widely quoted model for P_{ij} is the Rasch model and the Rasch conditional model (T_{ij}) is:

$$T_{ij} = \ln \left[\frac{P_{ij}(\theta)}{Q_{ij}(\theta)} \right] = -\delta_j \quad (1)$$

where δ_j is the item parameter. Throughout the remaining part of this paper, we denote $\hat{\delta}_1$, $\hat{\delta}_F$ or $\hat{\delta}_M$ as the estimated item parameters for the full, female and male sample. Given two estimated item difficulty parameters, $\hat{\delta}_1$ and $\hat{\delta}_M$ and, define $\hat{q}_M = (\hat{\delta}_M - \hat{\delta}_1)$ and $\hat{q}_F = (\hat{\delta}_F - \hat{\delta}_1)$ where \hat{q}_M is the estimated differences between the full difficulty parameter, and male difficulty parameter; \hat{q}_F is the estimated differences between the full

difficulty parameter and female difficulty parameter. Estimating the difficulty parameter in the two subgroups separately amounts to the estimating of $2k$ parameters - that is, the estimated difficulty parameter in the first group is different from the estimated difficulty parameter in the other group (Fisher & Molenaar, 1995) and $(k = n-1)$ is the length of the estimated difficulty parameter. Using the above illustration, the Hausman test for female students is:

$$H_F = \hat{q}_F^T \left[\hat{V} ar(\hat{\delta}_F) - \hat{V} ar(\hat{\delta}_1) \right]^{-1} \hat{q}_F \quad (2)$$

The complementary test for male students would be to compare $\hat{\delta}_1$ and $\hat{\delta}_M$ and the Hausman test for that will be

$$H_M = \hat{q}_M^T \left[\hat{V} ar(\hat{\delta}_M) - \hat{V} ar(\hat{\delta}_1) \right]^{-1} \hat{q}_M \quad (3)$$

The Hausman test equations in (2) and (3) have asymptotically a null $\chi^2(k)$ distribution where k ($k=n-1$) is the length of the estimated difficulty parameter $\hat{\delta}_1$. $\hat{V} ar(\hat{\delta}_1)$, $\hat{V} ar(\hat{\delta}_F)$ and $\hat{V} ar(\hat{\delta}_M)$ are the asymptotic variance of $\hat{\delta}_1$, $\hat{\delta}_F$ and $\hat{\delta}_M$. \hat{q}_M^T and \hat{q}_F^T are the transpose of \hat{q}_M and \hat{q}_F .

Procedures for bootstrapping the Hausman test in Rasch

The bootstrap estimates the sampling distribution of a statistic by iteratively resampling items with replacement from the observed data. One possible advantage of the bootstrap method is that it is a shortcut to create a proxy population size through large replications and can handle virtually any statistic (Efron & Tibshirani, 1993). Unfortunately and in most cases, the bootstrap with replacement method does not resemble the actual estimates. According to Schervish, (1994), the degree to which the replacement method is successful depends on the resemblance of the actual estimate. In bootstrap or simulation method, the accuracy of the forecast statistics is related to the number of replicates in the sample. If the number of replicates is increased, the sampling distribution will become

narrower and the confidence interval around the true value of the sampled statistics will become tighter.

In the remaining part of this section, we will present the procedures for using the bootstrap data to estimate the Hausman test in Rasch. Using a bootstrap method to estimate the Hausman test in Rasch for N persons ($i = 1, \dots, N$) and n items ($j = 1, \dots, n$) consists of the following basic steps:

1. Draw a sample with replacement from a sample data X_{ij} where, $X_{ij} = (X_{i1}, X_{i2}, \dots, X_{in})$ is the original data and X_{ij} is an order $N \times n$ matrix of persons by items, Denote the bootstrap sample as $X_{i1}^{*b}, \dots, X_{in}^{*b}$, where $b=1, \dots, B$ and $B=1,000$. The subgroup replicates is proportional to the size of the sample data.
2. Use the expression in equation (1) to estimate the item parameter.
3. Then the bootstrap method for estimating the Hausman test for Rasch in the male (M) sample is:

$$H_M^* = \hat{q}_M^{*T} V(S_M^*)^{-1} \hat{q}_M^* \quad (4)$$

where $V(S_M^*) = \left[\hat{V}ar(\hat{\delta}_M^*) - \hat{V}ar(\hat{\delta}_1^*) \right]$ and \hat{q}_M^{*T} is the transpose of \hat{q}_M^* . For clarity, star represents the values estimated from the bootstrap sample. The expression in equation (4) can be similarly represented for the female sample.

Method

Participants

The participants in the study comprised 200 students (male=120, female = 80) ranging in age from 15 to 19 years (mean = 16.5years, SD = 1.8) from the Alpha group of schools, in the Edo State - Nigeria.

Instrument

The mathematics test used in the study is a standardised one-hour, high-stakes, university entrance educational assessment that consisted of 10 dichotomously scored questions. The test was specifically constructed for students in Senior Secondary School class 3 (equivalent of year 12 in most developed countries) who were preparing for the West African Examination Council. The West African Examination Council question papers have been shown to have a high validity and reliability (see, Akubuiro & Joshua, 2004). The test contained five statistical questions (e.g., mean, median, mode, standard deviation and variance) and five questions based on general mathematics questions (e.g., series and sequences, trigonometry, basic calculus and geometry).

Analysis

Firstly, *RUMM2010* (Andrich, Sheridan & Luo, 2004) was used to produce the item characteristic curve and other fit statistics in Rasch. The item characteristic curves were tested by a chi-square statistic to detect item misfit. According to Hambleton (1993), plotting the observed versus expected score distribution allows for a visual representation of the fit between the two distribution.

We eliminated the person parameters and applied the conditional logistic method to items on the equal interval logit scale by using a log-linear formula in (1). We then used Rasch analysis to calibrate the items on a linear scale on the basis of subgroup by following a similar procedure employed by Andersen (1973a). Then samples of 1,000 examinees were generated from the full sample. For the two subgroups of males and females, the number of replicates was proportional to the size of the sample data. For example, 600

examinees were generated with replacement or simulated from male students which comprises of 120 samples while 400 examinees were generated with replacement or simulated from female students which comprises of 80 samples (see, Table 1). The bootstrap and simulation method for estimating the Hausman test in Rasch was estimated using R statistical computing (available at www.r-project.org). The R statistical program for the simulation method is available from the authors upon request.

Results

The sample comprised largely disadvantaged students from low socio-economical backgrounds (68% had an annual family income less than \$5,000; only 9% of the sample had a father with secondary or higher education; and only 2.5% had a mother with secondary or higher education). Of the 80 female student used in the analysis, about 61.3% are rich family while the rest are either poor or middle family where as for boys 72.5% are poor family and 10% are from rich family (see Table 1).

Table 1
Demographic characteristics of subjects by gender

Demographic information	Girls (n=80)	Boys (n=120)	Combine (N=200)
Age (Mean± sd)	(16.2± 1.8)	(16.6± 1.8)	(16.5±1.8)
Family income			
Poor (%)	8/80 (10.0)	87/120 (72.5)	136/200 (68.0)
Middle (%)	23/80 (28.7)	21/120 (17.5)	44/200 (22.0)
Rich (%)	49/80 (61.3)	12/120 (10.0)	20/200 (10.0)
Father's Occupation			
Civil service (%)	6/80 (7.5)	12/120 (10.0)	18/200 (9.0)
Small scale farmer (%)	48/80 (60.0)	55/120 (46.8)	103/200 (51.5)
Petty trader (%)	26/80 (32.5)	53/120 (44.2)	79/200 (39.5)
Mother's Occupation			
Civil service (%)	1/80 (1.2)	4/120 (3.3)	5/200 (2.5)
Small scale farmer (%)	16/80 (20.0)	30/120 (25.0)	46/200 (23.0)
Petty trader (%)	63/80 (78.8)	86/120 (71.7)	149/200 (74.5)

Table 2 presents the mean and standard deviation of the two subgroups examined in the study. The item means for male students ranged between 0.16 – 0.28 while the means of female students ranged between 0.13 - 0.31.

Table 2
Means, standard deviation of male, female and combined for the 10 items

Items	Boys (N = 120)		Girls (N = 80)		Combined (N = 200)	
	Mean	SD	Mean	SD	Mean	SD
Item 1	0.28	0.45	0.31	0.47	0.29	0.45
Item 2	0.20	0.40	0.14	0.35	0.18	0.38
Item 3	0.21	0.41	0.15	0.36	0.19	0.39
Item 4	0.16	0.37	0.14	0.35	0.15	0.36
Item 5	0.21	0.41	0.18	0.38	0.20	0.40
Item 6	0.23	0.42	0.18	0.38	0.21	0.40
Item 7	0.25	0.43	0.15	0.36	0.21	0.41
Item 8	0.18	0.39	0.23	0.42	0.20	0.40
Item 9	0.23	0.42	0.15	0.36	0.20	0.40
Item 10	0.22	0.41	0.13	0.33	0.18	0.39

Note: The mean is the proportion correct

The full data set consisting of responses from all participants in the study (N=200) was used for this RUMM analysis. From the RUMM analysis, all items except item 6 misfit with statistically significant overall χ^2 test-of-fit values ($p < 0.001$). The item characteristic curves for the original sample on the ten items are shown in Figure 1. In Figure 1, there is no close conformity of the data with the model. The item-trait interaction statistics, which identify the degree of the overall fit of the index to the Rasch model was significant ($\chi^2 = 168.69$, $df = 20$, $p < 0.0001$). This indicates that the full data are deviating significantly from the model. The person separation index and the likelihood ratio test was 0.60 and 0.61 respectively. The item location (logits) ranges from -0.01 to 0.35, the range of residual test of fit statistics was 0.00 to 0.18 with a standard deviation of 0.23 to 2.62 while the range of person fit was -1.74 to 0.35 with a standard deviation of 1.22 to 0.63.

The RUMM analysis from the simulated data indicates that all the items have misfit with statistical significant overall χ^2 test-of-fit values are $p < 0.001$. The degree of the overall fit of the index to the Rasch model was not significant ($\chi^2 = 1038.8$, $df = 50$, $p < 1.00$). The person separation index and the likelihood ratio test were both 0.57. The range of the residual test of fit statistics was 0.00 to 0.48 with a standard deviation of 0.22 to 5.62 and

the range of person fit was -1.72 to 0.35 with standard deviation of 1.18 to 0.62. For the bootstrap RUMM analysis (see Figure 3), all items fit with statistical overall are $p > 0.05$.

Overall, the original data fit of the items to the Rasch model is poor and further analysis to determine the model statistics was carried out by considering the bootstrap and simulation method for estimating the Hausman test in Rasch.

Figure 1. Item characteristic curves for the original data.

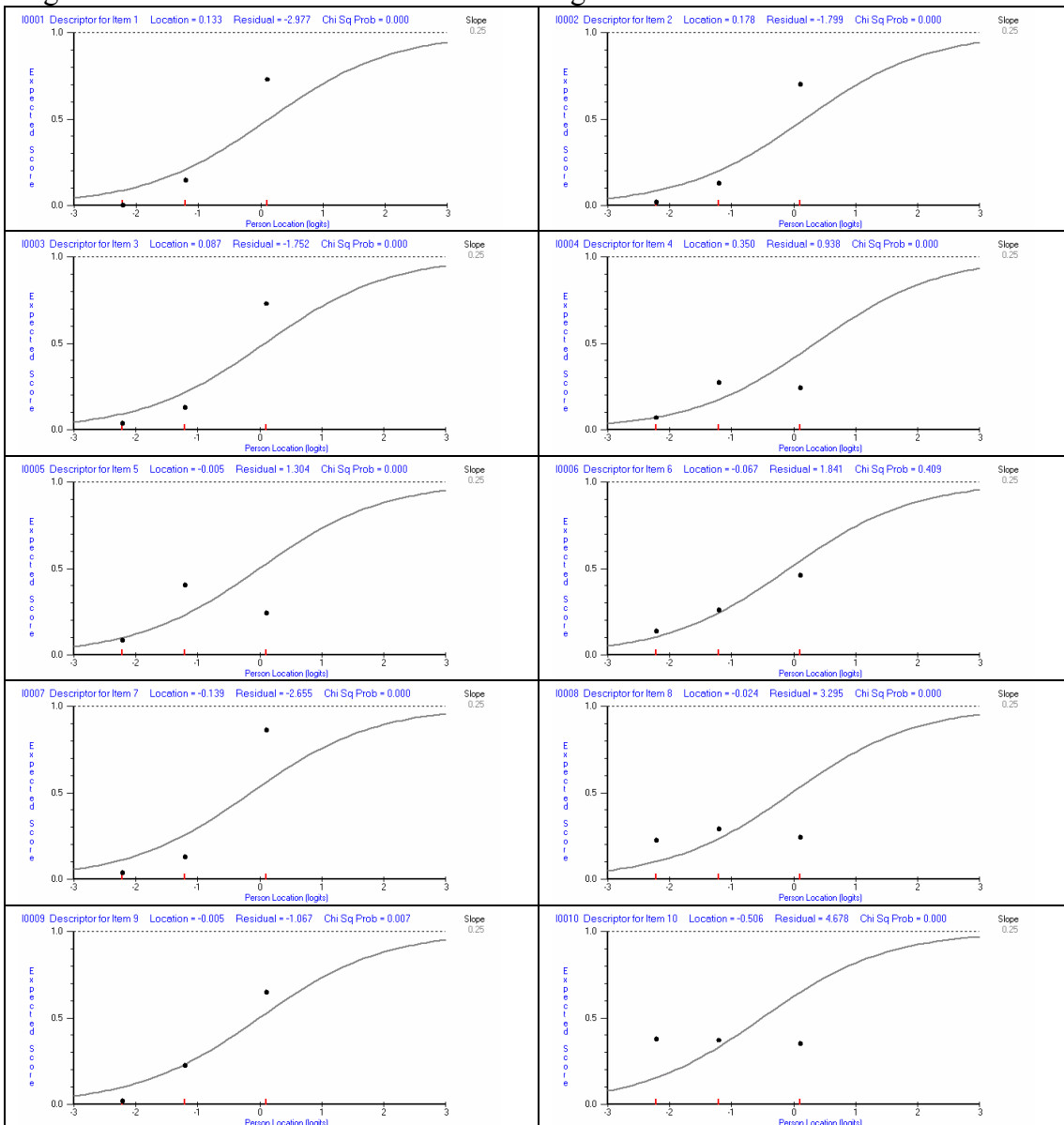


Figure 2. Item characteristic curves for the simulated data.

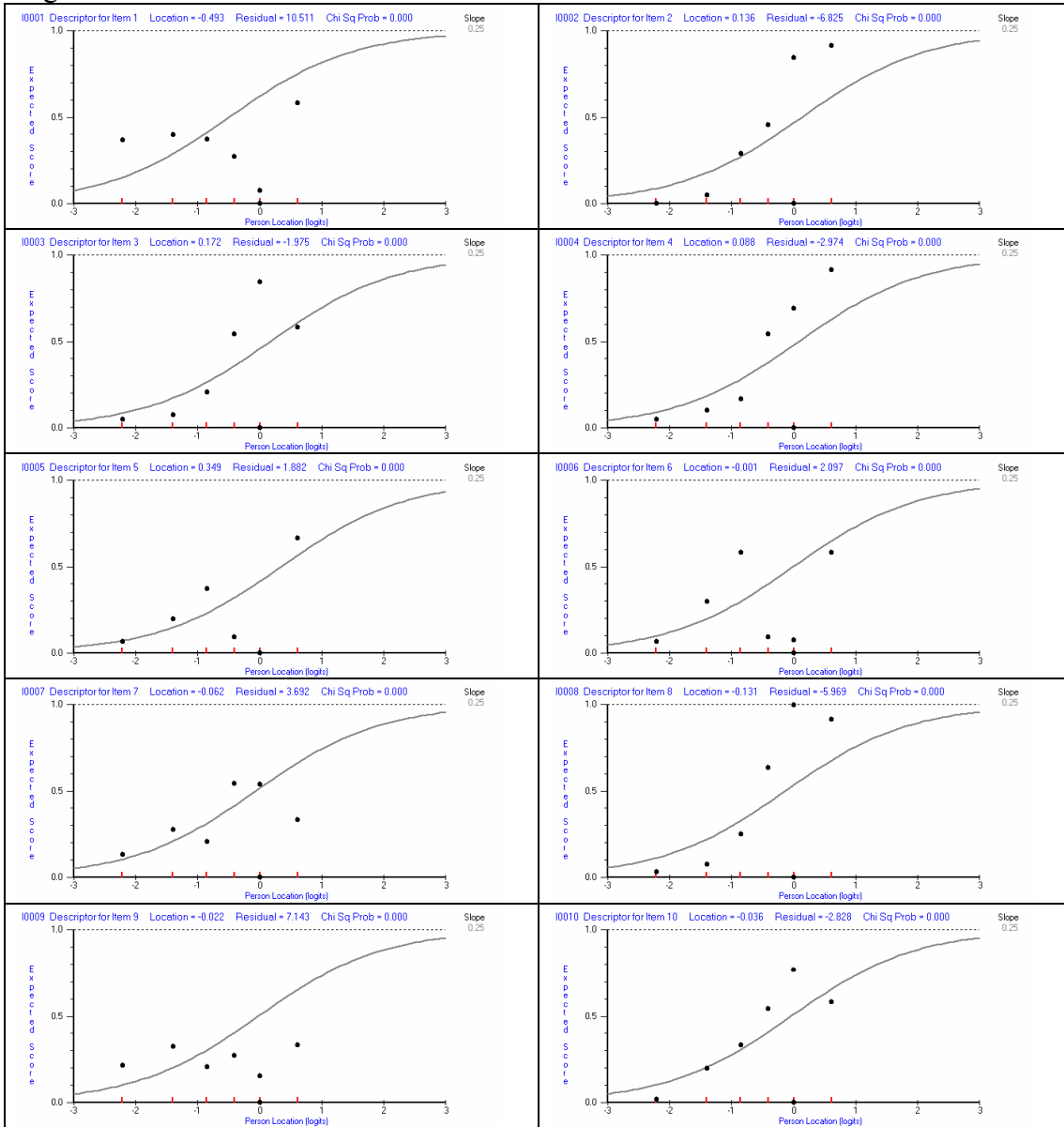
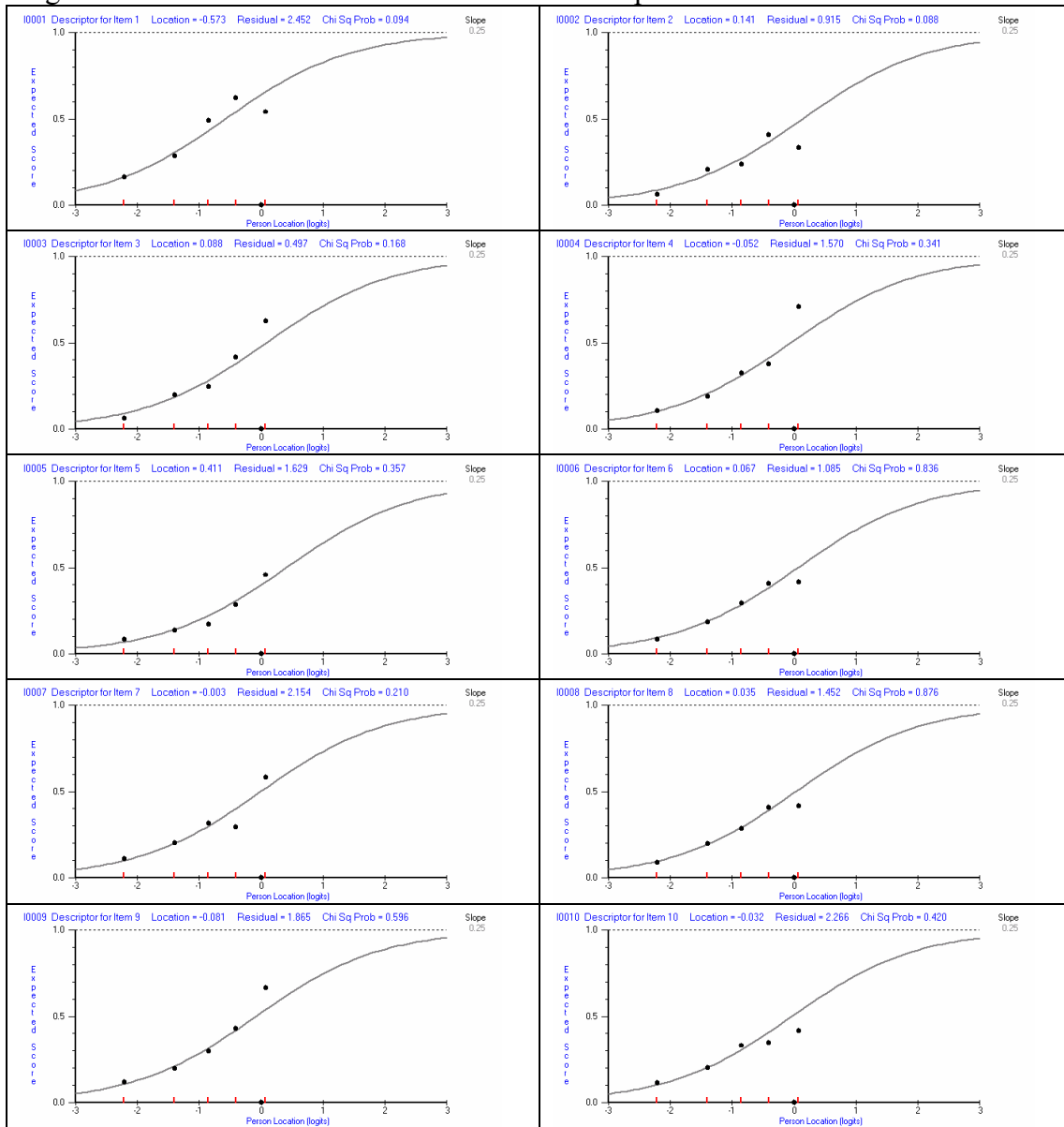


Figure 3. Item characteristic curves for the bootstrap data.



Tables 3, 4 and 5 report the parametric, simulation and bootstrap (with replacement) item difficulty parameters for male and female students. These are reported as restricted, the full estimates (unrestricted), the standard error, and the differences between the restricted and the unrestricted item difficulty calibrations. The unrestricted item difficulty calibrations in Rasch actually represent the values for male and females combined. They are cited in column 2 and 6 to show how the differences (\hat{q}_M or \hat{q}_F) was obtained. In Table 3, 4 and 5, the first item (item 1), with associated beta (1) = 0 is the reference level.

This is because, the conditional logit function in r-statistical programming is not able to define a reference level if item 1 is added.

The parametric approach did not adequately model the item difficulty parameter in Rasch. Looking at the chi-square (χ^2), the items were poorly fitted; for males ($\chi^2(9) = 7.21, p=0.61$) and for females ($\chi^2(9) = 7.57, p=0.58$). The simulation and bootstrap methods using the Hausman tests were statistically significant (see the last two rows of Tables 4 and 5) and were better in modelling the responses than the parametric method. For instance, the χ^2 value for males using the simulation method was ($\chi^2(9) = 36.07, p < 0.0001$) and the bootstrap method the ($\chi^2(9) = 88.33, p < 0.0001$). A similar significant interpretation applied to females. However, the reader will notice that the bootstrap item difficulties are generally lower than those of the simulation method but the standard errors are the same for the two methods.

Table 3
Item difficulty calibrations for the Hausman test in Rasch

Items	Item difficulty calibration for males				Item difficulty calibration for females			
	Restricted ($\hat{\delta}_M$)	Unrestricted ($\hat{\delta}_i$)	Difference (\hat{q}_m)	S.E = sqrt (diag($\hat{\delta}_M - \hat{\delta}_i$))	Restricted ($\hat{\delta}_F$)	Unrestricted ($\hat{\delta}_i$)	Difference (\hat{q}_f)	S.E = sqrt (diag($\hat{\delta}_F - \hat{\delta}_i$))
Item 2	0.37	0.71	-0.34	0.20	1.27	0.71	0.56	0.35
Item 3	0.49	0.75	-0.26	0.20	1.15	0.75	0.40	0.34
Item 4	0.43	0.67	-0.25	0.20	1.04	0.67	0.37	0.33
Item 5	0.81	0.95	-0.15	0.22	1.15	0.95	0.19	0.33
Item 6	0.43	0.60	-0.17	0.21	0.84	0.60	0.24	0.31
Item 7	0.31	0.53	-0.21	0.20	0.84	0.53	0.31	0.31
Item 8	0.15	0.49	-0.34	0.20	1.04	0.49	0.54	0.33
Item 9	0.61	0.56	0.05	0.22	0.49	0.56	-0.07	0.28
Item 10	0.26	0.56	-0.30	0.20	1.04	0.56	0.47	0.33
χ^2			7.21				7.57	
P value			0.61				0.58	

Note: Item 1 is not included because it is the reference level

Table 4
Item difficulty calibrations using the simulation method for the Hausman test in Rasch

Items	Item difficulty calibration for male				Item difficulty calibration for female			
	Restricted ($\hat{\delta}_M$)	Unrestricted ($\hat{\delta}_1$)	Difference (\hat{q}_u)	S.E = sqrt (diag($\hat{\delta}_M$ - $\hat{\delta}_1$))	Restricted ($\hat{\delta}_F$)	Unrestricted ($\hat{\delta}_1$)	Difference (\hat{q}_r)	S.E = sqrt (diag($\hat{\delta}_F$ - $\hat{\delta}_1$))
Item 2	0.37	0.71	-0.34	0.09	1.27	0.71	0.56	0.16
Item 3	0.49	0.75	-0.26	0.09	1.15	0.75	0.40	0.15
Item 4	0.43	0.67	-0.25	0.09	1.04	0.67	0.37	0.15
Item 5	0.81	0.95	-0.15	0.10	1.15	0.95	0.19	0.15
Item 6	0.43	0.60	-0.17	0.09	0.84	0.60	0.24	0.14
Item 7	0.31	0.53	-0.21	0.09	0.84	0.53	0.31	0.14
Item 8	0.15	0.49	-0.34	0.09	1.04	0.49	0.54	0.15
Item 9	0.61	0.56	0.05	0.10	0.49	0.56	-0.07	0.13
Item 10	0.26	0.56	-0.30	0.09	1.04	0.56	0.47	0.15
χ^2	36.07				37.87			
P value	< 0.0001				< 0.0001			

Note: Item 1 is the reference level

Table 5
Item difficulty calibrations using the bootstrap method for the Hausman test in Rasch

Items	Item difficulty calibration for male				Item difficulty calibration for female			
	Restricted ($\hat{\delta}_M^*$)	Unrestricted ($\hat{\delta}_1^*$)	Difference (\hat{q}_u^*)	S.E = sqrt (diag($\hat{\delta}_M^*$ - $\hat{\delta}_1^*$))	Restricted ($\hat{\delta}_F^*$)	Unrestricted ($\hat{\delta}_1^*$)	Difference (\hat{q}_r^*)	S.E = sqrt (diag($\hat{\delta}_F^*$ - $\hat{\delta}_1^*$))
Item 2	0.20	0.71	-0.50	0.08	0.87	0.71	0.16	0.14
Item 3	0.29	0.65	-0.36	0.09	0.76	0.65	0.10	0.14
Item 4	0.30	0.51	-0.21	0.09	1.02	0.51	0.51	0.15
Item 5	0.41	0.98	-0.57	0.08	0.85	0.98	-0.13	0.14
Item 6	0.12	0.63	-0.51	0.08	0.62	0.63	-0.02	0.13
Item 7	0.10	0.56	-0.46	0.08	0.62	0.56	0.05	0.13
Item 8	0.08	0.60	-0.53	0.08	0.60	0.60	0.00	0.13
Item 9	0.33	0.49	-0.16	0.09	0.23	0.49	-0.26	0.12
Item 10	0.10	0.53	-0.43	0.08	1.09	0.53	0.56	0.16
χ^2	88.33				45.82			
P value	< 0.0001				< 0.0001			

Note: Item 1 is the reference level

Tables 4 and 5 present the variation of difficulty estimates for the simulation and bootstrap methods. As expected the item difficulty values for the simulation method were identical with those of the parametric method because the replicate procedure increased sample size but did not alter item difficulty estimates. Simulation however altered the standard error (see columns 5 and 9 in Table 4). The difficulty estimate values obtained using bootstrap method were substantially lower than that of the simulation method. This is because the bootstrap method yields different values as the data process is repeated and the bootstrap method makes no assumptions about the population from which items and persons are sampled. In spite of the bootstrap method with replacement, the results for the parametric and bootstrap methods suggested that the female students found the mathematics items much difficult than the male students. This reflected the socioeconomically pattern of attendance in secondary schooling, where there is a tendency for secondary education to be limited for girls from richer families.

Discussion

An analysis of goodness of fit, whether the purpose is overall evaluation or isolation of bad data, needs to be guided by conceptions of the reasonable alternatives to the latent trait model under consideration (Traub & Wolfe, 1981). In this study, we recommended the creation of a proxy population size through replication and the Hausman test method to determine the unrestricted versus restricted of the replicated response processes in different population of sexes.

This present study investigated items that were poorly fitted and critically examined the bootstrap method for estimating the Hausman test for Rasch. The bootstrap method was beneficial for the dichotomous because the data fit the model better than the parametric (see χ^2 values in Table 3 and 5). The Hausman test can also be used to compare the difficulty estimates obtained from the full sample with the difficulty estimates obtained from the restricted male and female samples. In most econometric testing, the Hausman test is usually interpreted as the nonparametric Likelihood Ratio test (see Wong, 1996).

The Rasch item difficulty calibration using the bootstrap technique was used to correct for bias and to deal with such poorly fitting items. If an item or all items do not fit the Rasch model well and there are other substantive reasons for retaining them, then there is no need to delete the item(s) or stop calibrating item difficulty in Rasch (Kolen & Whitney, 1981; Rentz & Bashaw, 1975; Wright & Stone, 1979); rather, one may consider the bootstrap method as an alternative because when it comes to the situation where the model fits the data. Georg Rasch is also cited as saying, “That the model is not true is certainly correct, no models are” (see, Hambleton et al., 1992). To discard items or persons that are inconsistent with the model is not always defensible and Hambleton et al. (1992) made the further point that curriculum specialist cannot be asked to narrow their test content for the sake of psychometric models. However, the chi-square fit statistics output are questionable (see, Traub & Wolfe, 1981; Keeves, Johnson & Afrassa, 2000). In particular, the asymptotic properties of the tests of chi-square fit statistics cannot be determined mathematically (For further discussion, see Van den Wollenberg, 1979).

This study again highlighted differences in item difficulties between the parametric and bootstrap method for estimating the Rasch analysis especially when items did not fit the Rasch model well. Overall, the bootstrap estimates of difficulty were lower but the item parameter fitted the Rasch model better. Based on the chi-square value, the bootstrap estimates were better than the simulation estimates and better than the parametric estimates in the case of poorly fitting items.

Douglas and Cohen (2001) suggested a semi-parametric item response method which required splitting items into two parts. Those items that fit well were estimated using parametric item response theory methods while those that did not fit well were estimated using a nonparametric method. The claim was made by Wright and Stone, (1979), Rentz and Bashaw, (1975), and Kolen and Whitney, (1981) that items that did not fit the Rasch model well should be deleted before item calibration in Rasch can be carried. Previous literature has failed to show that the application of fit statistics in Rasch is subject to some argument because of the fundamental incompatibility between the data and model.

For example, if a calibration is small, the item parameter estimate and the ensuing statistical application will be inaccurate (Traub & Wolfe, 1981).

Second, the simulation and the parametric method for estimating the Hausman test for Rasch showed that item difficulty was the same for males and females. This finding again suggests that the Hausman test and the Likelihood Ratio have the same test of hypotheses (see, Andersen, 1973a,b) because they have the same asymptotic χ^2 -distribution; these are tests that focus on the assumptions of sufficiency and separability in Rasch and can be used in connection with any partitioning of the data set (Fisher & Molenaar, 1995; Weesie, 1999). On occasions the Hausman statistics may be negative. Hausman and McFadden (1984) argued that in such cases the statistic should be taken to be insignificant.

The main limitation of this study is that the bootstrap with replacement method tends to confirm the overoptimistic assessment of goodness of fit produced by the Hausman specification model (Schervish, 1994) and the bootstrap item difficulty estimates produce in this study does not resembles the parametric or simulation estimates (see Table 2, 3 & 4) which makes it difficult to take the bootstrap with replacement estimates seriously as a statistical tool (Schervish, 1994). Despite the criticism, the bootstrap method in statistics is the most efficient way for estimating the sampling distribution by resampling with replacement from the original data.

However, based on our results, it seems reasonable to recommend the bootstrap method for estimating the Hausman test in Rasch for existing Rasch packages because the bootstrap and simulation method were statistically better than the parametric method and the Hausman specification test avoided the problem of applying the maximum likelihood theory to joint estimation employed by Andersen (1973 a,b). For instance, there should be an existing package if practitioners desire to estimate the item parameter goodness of fit for small sample sizes, non-representative study samples and when the item Infit and Outfit fit statistics in Rasch are poorly fitting for the original data. This will reduce the time required to write the programming techniques suggested in this study.

Acknowledgments

The assistance of Mr. Taiwo Uyi Aiworo, Principal the Alpha group of schools, in Edo State-Nigeria for helping in data collection is gratefully acknowledged together with the support of Ms Sonia Freeman from the Centre for Clinical Epidemiology and Biostatistics and the helpful comments of two anonymous reviewers.

References

- Adams, R. J. and Khoo, S.T. (1993). *Quest-The interactive test analysis system*. Hawthorn, Victoria: ACER.
- Akubuiro, I.M and Joshua, M.T (2004). Self-concept, Attitude and Achievement of Secondary School Students in Science in Southern Cross River State, Nigeria. Available at <http://www2.ncsu.edu/ncsu/aern/seksiens.html>.
- Andersen, E.B. (1973a). A goodness of fit test for the Rasch model. *Psychometrika*, 38, 123-140.
- Andersen, E. B. (1973b). Conditional inference and multiple choice questionnaires. *British Journal of Mathematics and Statistical Psychology*, 26, 31 – 44.
- Andrich, D., Sheridan, B.E & Luo, G. (2004). *Rasch unidimensional measurement Models (RUMM): A windows based computer program*. Perth: Murdoch University.
- Douglas, J. & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, 25, 234 – 243.
- Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *Annals of Statistics*, 7, 1-16.
- Efron, B & Tibshirani, J.C (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.
- Fisher, G. H. & Molenaar, I.W. (1995). *Rasch models. foundations, recent developments and applications*. New York: Springer-Verlag.
- Hambleton, R.K. (1993). Principles and selected application of item response theory. In R.L. Linn (E.d.) *Educational measurement*. (3rd ed.) (pp.147-200). Phoenix: The Orxy Press.
- Hambleton, R., & others. (1992). Hambleton's 9 theses. *Rasch Measurement Transactions*, 6(2), 215 [<http://www.rasch.org/rmt/rmt62.htm>].
- Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6):1251–1271.
- Hausman, J. & McFadden, D. (1984). Specification test for the multinomial logit model, *Transportation Research*, 15B, 345 - 360.
- Junker, B.W. & Sijtsma, K. (2001a). Nonparametric item response theory in action. An overview of the special issue. *Applied Psychological Measurement*, 25, 211 – 250.
- Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement*, 2(4), 389-423.
- Keeves, J.P., Johnson, T.G., & Afrassa, T.M. (2000). Errors: What are they and how significant are they? *International Education Journal* 1, (3), 164-180
<http://www.flinders.edu.au/education/i.e>.
- Kolen, M.J. & Whitney, D. R. (1981). *Comparison of four procedures for equating the tests of general educational development*. Paper presented at the annual meeting of three American Educational research association. Los Angeles, California.
- Linacre J.M. & Wright B.D. (1994) Chi-Square Fit Statistics. *Rasch Measurement Transactions*, 8:2, 360

- Molenaar, I. W. (1995). Nonparametric methods for polytomous responses. In W. J. van der Linden & R. K. Hambleton (Eds.) *Handbook of modern item response theory*, pp. 369 – 380. New York: Springer.
- Molenaar, I. W. (2001). Thirty years of nonparametric item response theory. *Applied Psychological Measurement*, 25, 295 – 299.
- Pfanzagl, J. (1994). On item parameter estimation in certain latent trait models. In G.H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology. Psychometrics and methodology*, pp. 249-263. NY: Springer-Verlag.
- Ramsay, J.O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56, 611 – 630.
- Rasch G. (1960/1980). *Probabilistic models for some intelligence and attainment test*. Chicago: University of Chicago Press. (Originally published by The Danish Institute for Educational Research, Copenhagen, 1960).
- Rentz, R.R. & Bashaw, W. L. (1975). Equating Reading tests with the Rasch model, *Vol. 1 Final report*. Athens, Georgia: University of Georgia: Educational Research Laboratory, College of Education.
- Robins, J.M., van der Vaart, A.W., and Ventura, V. (2000). The Asymptotic Distribution of P-Values in Composite Null Models. *Journal of the American Statistical Association*, 95 (452), 1143-1156.
- Schervish, M.J. (1994). Bootstrap: More than a Slab in the Dark. *Statistical Science*, 9(3), 408-410.
- StataCorp. *Stata Statistical Software, Release 7.0*. In. College Station, TX: Stata Corporation; 2001.
- Stout, W. F. (2001). Nonparametric item response theory. A maturing and applicable measurement modelling approach. *Applied Psychological Measurement*, 25, 300 – 306.
- Weesie, J. (1999). *The Rasch model in STATA*. STATA statistical software 7.0 : STATA Corporation.
- Traub, R.E and Wolfe, R.G (1981). Latent Trait Theories and the Assessment of Educational Achievement. *Review of Research in Education*, 9, 377- 435.
- Van den Wollenberg, A. L. *The Rasch model and time limit test: An application and some theoretical contributions*. Druk: Strichting Studentenpers Nijmegen, 1979.
- Wong, K. (1996) Bootstrapping Hausman's exogeneity test, *Economics letters*, 53 139-143.
- Wright, B.D. & Panchapakesan, N. A (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 29, 23-48.
- Wright, B.D. & Stone, M.H. (1979). *Best test Design. Rasch measurement*. Chicago MESA Press.